Optimal design of experiments with application to the inference of traffic matrices in large networks: second order cone programming and submodularity

Guillaume Sagnol

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUT OF TECHNOLOGY

MINES ParisTech

École doctorale n°432 : Sciences des Métiers de l'Ingénieur

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

l'École nationale supérieure des mines de Paris

Spécialité « Informatique temps réel, robotique, automatique »

présentée et soutenue publiquement par

Guillaume SAGNOL

le 13 décembre 2010

Plans d'expériences optimaux et application à l'estimation des matrices de trafic dans les grands réseaux

Programmation conique du second ordre et Sous-modularité

Directeurs de thèse : **Stéphane GAUBERT**
Yves ROUCHALEAU
Co-encadrement de la thèse : **Mustapha BOUHTOU**

Jury

M. Walid BEN-AMEUR, Professeur, TELECOM&Management SudParis	Raporteur
M. Anatoly ZHIGLJAVSKY, Professeur, School of Mathematics, Cardiff University	Rapporteur
M. Jean-Baptiste HIRIART-URRUTY, Professeur, Université Paul Sabatier, Toulouse	Examinateur
M. Michel MINOUX, Professeur, Laboratoire d'informatique de Paris 6	Examinateur
M. Mustapha BOUHTOU, Responsable de projet, Orange Labs R&D	Co-encadrant
M. Stéphane GAUBERT, Directeur de recherche, INRIA Saclay & École Polytechnique	Directeur
M. Yves ROUCHALEAU, Professeur, Centre de Mathématiques Appliquées, Mines Paristech	Co-directeur

MINES ParisTech
Nom de l'Unité de recherche
Adresse de l'Unité recherche

T H È S E

THÈSE

pour obtenir le grade de

DOCTEUR DE L'ÉCOLE NATIONALE SUPÉRIEURE DES MINES DE PARIS

Spécialité **Informatique temps réel, robotique, automatique**

Présentée par

Guillaume SAGNOL

Plans d'expériences optimaux et application à l'estimation des matrices de trafic dans les grands réseaux

Programmation conique du second ordre et Sous-modularité

Optimal design of experiments with application to the inference of traffic matrices in large networks

Second order cone programming and Submodularity

Soutenue le 13 décembre 2010 devant un Jury composé de

M. Walid BEN-AMEUR	Raporteur
M. Anatoly ZHIGLJAVSKY	Rapporteur
M. Jean-Baptiste HIRIART-URRUTY	Examinateur
M. Michel MINOUX	Examinateur
M. Mustapha BOUHTOU	Co-encadrant
M. Stéphane GAUBERT	Directeur
M. Yves ROUCHALEAU	Co-directeur

Table des matières

Résumé

Les fournisseurs d'accès Internet souhaitent avoir une bonne connaissance du trafic traversant leur réseau, pour de nombreuses opérations contribuant à la bonne gestion du trafic et à la maintenance du réseau. Une partie essentielle de l'information dont ils ont besoin pour ces tâches est la *matrice de trafic*, qui indique les volumes de trafic pour chaque paire origine-destination du réseau pendant un laps de temps donné, c'est à dire le nombre d'octets ayant transité depuis chaque nœud d'entrée vers chaque nœud de sortie pendant la période considérée. L'importance des opérations d'ingénierie du trafic s'appuyant sur la donnée de cette matrice ne cesse d'augmenter, puisque le trafic à traiter augmente et se diversifie, devenant plus complexe d'année en année. Mais en pratique, il est très difficile d'obtenir des estimations précises des demandes de trafic en origine-destination. Contrairement à ce que l'intuition peut laisser croire, les mesures sur les réseaux sont : (i) souvent indisponibles au niveau de certains routeurs non instrumentés ; (ii) coûteuses ; (iii) susceptibles d'affecter la qualité de service. Les décision concernant l'emplacement des mesures à prendre, ainsi que leur taux d'échantillonage constituent donc un enjeu crucial.

Nous abordons le problème de l'optimisation des mesures dans les réseaux par une approche fondée sur la *théorie des plans d'expériences optimaux*. Cette théorie étudie comment allouer l'effort expérimental à un ensemble d'expériences disponibles, quand le but est de maximiser la qualité de l'estimation d'un *paramètre inconnu*. Si l'on considère chaque localisation possible du logiciel de mesure comme une *expérience*, et la matrice de trafic comme le *paramètre inconnu*, on obtient une formulation de type *plans d'expériences* de notre problème de télécommunications. Cependant, les algorithmes classiques en conception optimale d'expériences se révèlent inefficaces sur les grands réseaux. Par ailleurs, la difficulté est augmentée par le fait que chaque mesure peut fournir plusieurs observations simultanées des demandes de trafic (conception optimale *d'expériences multiréponses*).

Dans la première partie de cette thèse, nous développons une approche fondée sur l'*Optimisation Conique du Second Ordre* (SOCP), pour résoudre des problèmes de grande taille en conception optimale d'expériences multiréponses. Un avantage *clé* de notre approche est que le *solver PCSO* ne gère que des matrices creuses et de tailles modérées, tandis que les algorithmes classiques ont besoin de gérer de grandes matrices pleines pour résoudre les même instances. De plus, l'approche par PCSO permet une grande flexibilité dans la définition des contraintes sur les plans d'expériences. Le cœur de notre méthode est un théorème de réduction du rang en optimisation semi-définie, qui permet une description géométrique simple des plans d'expériences optimaux. Certains aspects combinatoires –qui apparaissent typiquement lorsque l'opérateur souhaite choisir un sous-ensemble de routeurs à instrumenter pour qu'ils puissent prendre des mesures– sont également étudiés. Grâce à des inégalités matricielles et à des techniques d'optimisation sous-modulaire, nous formulons des bornes sur la performance de l'algorithme glouton et de techniques d'arrondis.

L'application à l'inférence des matrices de trafic dans les réseaux de télécommunication fait l'objet de la seconde partie de ce manuscrit. Lorsque l'on dispose uniquement de mesures partielles sur le réseau, l'état de l'art est une méthode –dite *tomogravitaire*– qui comble les données manquantes en résolvant des problèmes de minimisation d'entropie. La qualité de l'estimation obtenue dépend toutefois grandement de la localisation et des taux d'échantillonage des mesures disponibles. Les expériences numériques présentées en première partie montrent que notre approche

par PCSO est la plus efficace pour le problème de conception $\boldsymbol{c}-$*optimale*, i.e. lorsque l'expérimentateur cherche à estimer une combinaison linéaire seulement des paramètres inconnus (dans notre cas, les demandes de trafic) ; nous développons donc une méthode –baptisée *plans successifs d'expériences* $\boldsymbol{c}-$*optimales*– dans laquelle on considère plusieurs combinaisons linéaires (tirées de façon aléatoire) des demandes de trafic. Notre approche est comparée aux précédentes, et évaluée sous de nombreux points de vue, par l'intermédiaire de simulations avec des données réelles. En particulier, nous traitons des instances pour lesquelles les approches précédentes étaient incapable de fournir une solution. Finalement, nous proposons de nouvelles directions pour les techniques d'estimation de la matrice de trafic dans un chapitre d'ouverture. Nous mettons en évidence la structure de petit rang des matrices de trafic, grâce à la théorie des matrices aléatoires et à des décompositions de tenseurs. Enfin, nous présentons l'esquisse préliminaire d'une approche tensorielle qui semble améliorer la méthode *tomogravitaire*.

Summary

Internet Service Providers (ISP) wish to have a good knowledge about the traffic which transit through their networks, for many traffic engineering and network planning tasks. An essential part of the required information is the *traffic matrix*, which contains the volumes of traffic for each origin-destination pair of the network during a given period of time, i.e. the number of bytes that has travelled from any entry node to any exit node. The importance of the networking operations relying on the traffic matrix is increasing as the traffic grows in volume and becomes more complex, but in practice, obtaining accurate estimations of the demands of traffic is a challenging issue. Contrarily to what intuition may suggest, network measurements are: (i) often not available everywhere; (ii) expensive; (iii) likely to affect the quality of service. It is thus a crucial issue to decide where network measurements should be performed, as well as their sampling rates.

We approach the problem of optimizing the network measurements by using the *theory of optimal experimental designs*. This theory studies indeed how to allocate the experimental effort to a set of available experiments, in order to maximize the quality of estimation of an *unknown parameter*. Thinking of each potential location of the measuring software as an *experiment*, and the traffic matrix as the *unknown parameter*, one obtains a nice *optimal experimental design* formulation of our telecommunications problem. However, the classic optimal experimental design algorithms are intractable on large scale networks, because very large matrices are involved. In addition, the difficulty is increased by the fact that each measurement yields several simultaneous observations of the unknown volumes of traffic (optimal design of *multiresponse experiments*).

In the first part of this thesis, we develop an approach relying on *Second Order Cone Programming* (SOCP) to solve large-scale, multiresponse optimal experimental design problems. An important advantage of our approach is that the SOCP solver handles sparse matrices of moderate size, while classic algorithms need store large full matrices to solve the same instances. Moreover, SOCP solvers allow one to define constraints on the experimental design with lots of flexibility. At the heart of our method is a rank reduction theorem in semidefinite programming, which allows a simple geometrical characterization of the optimal designs. Some combinatorial problems –which typically arise when an ISP wants to choose a subset of routers to upgrade, so that they will support a measuring software– are also studied. Thanks to matrix inequalities and submodular optimization techniques, we specify some lower bounds for the performance of greedy and rounding algorithms.

The application to the inference of the traffic matrix in telecommunication networks is the object of the second part of this manuscript. When partial measurements are available, the state of the art is the so-called *tomogravity method*, in which the lack of information is handled by solving entropy minimization problems. The quality of the obtained estimation nevertheless depends grandly of the localization and sampling rates of the available measurements. The numerical experiments presented in the first part show that our SOCP approach is most efficient for the $\boldsymbol{c}-$optimal design problem, i.e. when the experimenter wants to estimate only a linear combination of the unknown parameters (in our case, the traffic demands); we therefore develop a method – called *successive* $\boldsymbol{c}-$*optimal designs*– in which several randomly drawn linear combinations of the traffic demands are considered. This approach is compared to previous ones, and is fully evaluated by mean of simulations relying on real data. In particular, we handle some instances that were previously intractable. Finally, new directions for the techniques of estimation of the traffic matrix are considered in a perspectives chapter. By mean of the theory of random matrices and tensor decompositions, we evidence the low-rank structure of traffic matrices. The preliminary sketch of a tensorial approach, which seems to improve on the classic *tomogravity* method, is presented.

Remerciements

Mes premiers remerciements vont naturellement à mon directeur de thèse Stéphane Gaubert, dont le soutien a été inestimable pendant ces trois années. Son expertise technique, sa culture scientifique, ses qualités pédagogiques, ses conseils avisés, et surtout sa capacité à se rendre disponible malgré un emploi du temps très chargé ont grandement contribué à l'aboutissement de ce travail. Je le remercie encore chaleureusement pour s'être investi pleinement dans cette thèse, pour m'avoir encouragé et pour m'avoir donné confiance en moi. Je remercie également mon co-directeur Yves Rouchaleau pour son soutien et sa confiance en mon travail, ainsi que ses nombreux conseils.

Cette thèse a été financée par un contrat de recherche entre Orange Labs et l'INRIA [1]. Je remercie Mustapha Bouhtou en premier lieu pour avoir été à l'initiative de ce contrat, mais aussi pour sa participation active dans ce projet et ses efforts pour me fournir des données. Les discussions fructueuses que nous avons eues sont à l'origine de nombreux résultats de cette thèse.

J'exprime ma vive gratitude envers mes rapporteurs Messieurs Walid Ben-Ameur et Anatoly Zhigljavsky pour l'intérêt qu'ils ont montré pour mes travaux et le temps qu'ils m'ont consacré en écrivant leurs rapports. Je remercie les examinateurs Messieurs Jean-Baptiste Hiriart-Urruty et Michel Minoux de m'avoir fait l'honneur d'évaluer ce travail de thèse.

J'adresse mes remerciements à Paul van Dooren et Mariya Ishteva, qui m'ont aidé pour l'utilisation des tenseurs lors d'une visite à l'université catholique de Louvain. Je remercie également toutes les personnes que j'ai côtoyées quotidiennement pendant cette thèse. Merci notamment à Marianne, qui m'a souvent aidé par l'intermédiaire de Stéphane, et à mes collègues du bureau 2009 au CMAP, Meisam, Jean-Baptiste, Denis et Abdul pour leurs nombreux conseils et leur convivialité.

Merci enfin à ma famille et mes amis. Vous avez toujours été là pour me soutenir... et me rapeler que la vie ne se résout pas à un problème de matrices. Merci tout particulièrement à Marion qui m'a permis d'avancer régulièrement. Merci pour ta patience, pour avoir supporté mes périodes de rush, et pour m'avoir aidé, avec Augustin, à traverser les moments de doute en toute sérénité.

Acknowledgement

I thank prof. Anatoly Zhigljavsky again, in english, for the profound work he has been doing in writing his report, and in particular for having pointed out an example with a quadratic regression.

1. contrat de recherche CRE EB 257676

List of notation

$\mathbf{0}$	vector of all zeros (of the appropriate dimension), page 28
$\mathbf{1}$	vector of all ones (of the appropriate dimension), page 28
A	Routing matrix, page 146
$A(\boldsymbol{x})$	Observation matrix of size $l(\boldsymbol{x}) \times m$ for the experiment at $\boldsymbol{x}$, page 30
A_i	Observation matrix of size $l_i \times m$, for the i^{th} experiment (for the case $\mathcal{X} = [s]$), page 29 / for the i^{th} location of Netflow, page 185
$A \preceq B$	Notation for the Löwner ordering: $A \preceq B \iff B - A$ is positve semidefinite, page 29
$A \prec B$	Notation for the strict Löwner ordering: $A \prec B \iff B - A$ is positve definite, page 29
$A \odot B$	Hadamard (elementwise) product of A and B, page 106
$A \circledast B$	Khatri-Rao product of A and B, page 219
$A \otimes B$	Kronecker product of A and B, page 219
$\boldsymbol{a_x}$	Vector notation for $A(\boldsymbol{x})^T$, when $A(\boldsymbol{x})$ is a row vector ($i.e.\ l(\boldsymbol{x}) = 1$), page 40
$\mathcal{A}$	Aggregate observation matrix: $\mathcal{A} = [A_1^T, \ldots, A_s^T]^T$, page 47
$A(\xi)$	Aggregate observation matrix for the experiments in $\xi = \{\boldsymbol{x_k}, w_k\}$: $A(\xi) = [A(\boldsymbol{x_1})^T, \ldots, A(\boldsymbol{x_s})^T]^T$, page 31
ALS	Alternating Least Squares, page 219
AS	Autonomous System, page 154
$\boldsymbol{c_i}$ $(i \in [r])$	Columns of K, such that $\zeta_i = \boldsymbol{c_i}^T \boldsymbol{\theta}$, page 31
$\text{cone}(\mathcal{S})$	Conic hull of $\mathcal{S}$, page 29
$\text{conv}(\mathcal{S})$	Convex hull of $\mathcal{S}$, page 29
$[\![\mathcal{C};\ U, V, W]\!]$	Tucker decomposition of a tensor: $[\![\mathcal{C};\ U, V, W]\!] := \sum_{k_1,k_2,k_3} c_{k_1,k_2,k_3}\ \boldsymbol{u_{k_1}} \circ \boldsymbol{v_{k_2}} \circ \boldsymbol{w_{k_3}}$, page 220
$\text{Diag}(\cdot)$	Diagonal matrix defined by its diagonal entries, page 29
$\mathbf{diag}(M)$	Vector containing the diagonal elements of M, page 29
$\partial(\cdot)$	Boundary of a set, page 40
DSH	Decreasingly subhomogeneous, page 177
$\mathcal{D}_\beta$	Generalized Elfving set for $S_\beta-$optimality: (cf. Equation (5.24)), page 101
$\mathcal{E}$	Elfving Set: $\mathcal{E} = \text{conv}\left(\{\pm \boldsymbol{a_x},\ \boldsymbol{x} \in \mathcal{X}\}\right)$, page 40
$\overline{\mathcal{E}}$	Generalized Elfving set for multiresponse experiments: $\text{conv}\left(\{A_i^T \boldsymbol{\epsilon_i},\ i \in [s],\ \boldsymbol{\epsilon_i} \in \mathbb{R}^{l_i},\ \|\boldsymbol{\epsilon_i}\|_2 \leq 1\}\right)$, page 86
$\boldsymbol{e_i}$	vector of all zeros with a 1 in position i, page 48
ESCOD	Expected value of the Successive $\boldsymbol{c}-$Optimal Designs, page 194

$\|\cdot\|_F$	Frobenius norm, page 29
$\boldsymbol{I}$	Identity matrix (of the appropriate size), page 29
$\boldsymbol{I}_n$	Identity matrix (of size $n \times n$), page 29
iid	Independent and identically distributed, page 113
$\mathrm{Im}\, M$	Vector space generated by the columns of M : $\{\boldsymbol{x} : \exists \boldsymbol{y} : M\boldsymbol{y} = \boldsymbol{x}\}$, page 29
IPF	Iterative Proportional Fitting, page 177
ISP	Internet Service Provider, page 145
K	Matrix such that the quantity of interest is $\boldsymbol{\zeta} = K^T\boldsymbol{\theta}$, page 31
$\mathrm{Ker}\, M$	Nullspace of the matrix M : $\{\boldsymbol{x} : M\boldsymbol{x} = \boldsymbol{0}\}$, page 29
l	Number of observation per experiment, when we assume that $l(\boldsymbol{x})$ is constant over $\mathcal{X}$, page 29 / Number of links in the network, page 146
$l(\boldsymbol{x})$	Number of simultaneous observations collected by a measurement at $\boldsymbol{x}$, page 30
$\lambda_{\mathbf{max}}(M)$	Largest eigenvalue of the symmetric matrix M, page 56
$\lambda_{\mathbf{min}}(M)$	Smallest eigenvalue of the symmetric matrix M, page 56
LMI	Linear matrix inequality, page 58
m	Number of unknown parameters, page 29 / Number of OD pairs, page 146
$M^\dagger$	Moore-Penrose generalized inverse of M, page 29
M^-	A generalized inverse of M, i.e. any matrix G verifying $MGM = M$. This notation is used in expressions that do not depend on the choice of the generalized inverse G, page 29
mod	*modulo* operator, page 156
$M(\boldsymbol{x})$	Partial information matrix of the experiment $\boldsymbol{x}$: $M(\boldsymbol{x}) := A(\boldsymbol{x})^T A(\boldsymbol{x})$, page 32
$M(\xi)$	Information matrix of the design $\xi = \{\boldsymbol{x_k}, w_k\}$: $M(\xi) = \sum_{k\in[s]} w_k M(\boldsymbol{x_k})$, page 32
N	Number of $\boldsymbol{c}-$optimal designs used in the SCOD procedure, page 193
n	Number of allowed observations, page 29 / Number of nodes in a network, page 146
$[n]$	Notation for $\{1, \ldots, n\}$, page 28
$\|\cdot\|$	L_2-norm, page 28
$\|\cdot\|_p$	L_p-norm, page 28
OD	Origin-Destination, page 146
p	Exponent involved in the Kiefer's criterion Φ_p, page 34
PCA	Principal Components Analysis, page 156
PCOS	Plans $\boldsymbol{c}-$optimaux successifs, page 9
p.d.f.	Probability distribution function, page 215
Φ_p	Kiefer's criterion, page 34
$\propto$	Proportional to, page 49
$Q_K(\xi)$	$K-$information matrix. In the feasible case $\xi \in \Xi(K)$, $Q_K(\xi)$ is the inverse of the covariance matrix of the best linear unbiased estimator for $\zeta = K^T\boldsymbol{\theta}$: $Q_K(\xi) = (K^T M(\xi)^- K)^{-1}$, page 33
r	Number of quantities of interest (dimension of $\boldsymbol{\zeta}$), page 31
$\langle\cdot,\cdot\rangle$	Inner product on $\mathbb{S}_m$: $\langle A, B\rangle = \mathrm{trace}(A^T B)$, page 29
s	The number of support points of the design, page 30
$\mathcal{S}^\perp$	Orthogonal of the set $\mathcal{S}$: $\{\boldsymbol{x} : \forall \boldsymbol{v} \in \mathcal{S}, \boldsymbol{x}^T\boldsymbol{v} = 0\}$, page 29

$\mathbb{S}_m$	Space of symmetric $m \times m$ matrices, page 29
$\mathbb{S}_m^+$	Space of symmetric positive semidefinite $m \times m$ matrices, page 29
$\mathbb{S}_m^{++}$	Space of symmetric positive definite $m \times m$ matrices, page 29
$\mathcal{S}^n$	$n-$dimensional unit sphere of $\mathbb{R}^{n+1}$, in the Euclidean norm: $\{\boldsymbol{u} \in \mathbb{R}^{n+1} : \ \|\boldsymbol{u}\|_2 = 1\}$, page 197
$\mathcal{S}_p^n$	$n-$dimensional unit sphere of $\mathbb{R}^{n+1}$ in the ℓ_p norm: $\{\boldsymbol{u} \in \mathbb{R}^{n+1} : \ \|\boldsymbol{u}\|_p = 1\}$, page 220
SCOD	Successive $\boldsymbol{c}-$optimal designs, page 183
SDP	Semidefinite Program(ming), page 51
SNMP	Simple Network Management Protocol, page 146
SOCP	Second Order Cone Program(ming), page 61
$\mathrm{supp}(\xi)$	Set of all measurement points $\boldsymbol{x_i}$ associated with a positive weight w_i (for the design $\xi = \{\boldsymbol{x_i}, w_i\}$), page 30
SVD	Singular Value Decomposition, page 156
T	Number of time intervals in the global observation period, page 146
$.^T$	Transposition operation, page 28
$\boldsymbol{\theta}$	Vector of dimension m of the unknown parameters, page 29
TM	Traffic Matrix, page 145
$\boldsymbol{u} \circ \boldsymbol{v}$	Outer product: $(\boldsymbol{u} \circ \boldsymbol{v})_{ij} = u_i v_j$, page 217
$[\![\boldsymbol{\sigma}; U, V, W]\!]$	CP decomposition of a tensor: $[\![\boldsymbol{\sigma}; U, V, W]\!] := \sum_{k=1}^{r} \sigma_k \ \boldsymbol{u_k} \circ \boldsymbol{v_k} \circ \boldsymbol{w_k}$, page 218
$\boldsymbol{w}$	Vector of the experimental design variables, page 30
$\mathcal{X}$	Set of available experiments (experimental region), page 29 / Traffic tensor, page 223
$\boldsymbol{x}$	Measurement point in $\mathcal{X}$, page 29 / Vectorized traffic matrix (of size m), page 146
X	Dynamic traffic matrix of size $m \times T$, page 146
$\boldsymbol{x_t}$	Snapshot of the (vectorized) traffic matrix at time t (t^{th} column of X), page 146
$\xi = \{\boldsymbol{x_k}, w_k\}$	Design with experiments at $\boldsymbol{x_1}, \ldots, \boldsymbol{x_s}$, with respective weights $w_1, \ldots, w_s$, page 30
$\Xi(K)$	Feasibility cone for an observation matrix K, i.e. set of the designs ξ such that $\mathrm{Im}\, K \subset \mathrm{Im}\, A(\xi)$, page 31
$X_{(i)}$	i^{th} mode unfolding of the tensor $\mathcal{X}$, page 219
X_t	Snapshot (at time t) of a $n \times n$ Origin-Destination traffic matrix, page 211
Y^{SNMP}	Dynamic link count matrix (of size $l \times T$), page 146
$\boldsymbol{y^{\mathrm{SNMP}}}$	Vector of link counts (of dimension l) (SNMP measurements), page 146
$\boldsymbol{y_t^{\mathrm{SNMP}}}$	Link counts at time t (t^{th} column of Y^{SNMP}), page 146
$\boldsymbol{y}(\boldsymbol{x})$	Vector of observations at $\boldsymbol{x}$, page 29
$\boldsymbol{\zeta}$	Quantities of interest that the experimenter wants to estimate, page 31

Chapitre 1

Introduction (en Français)

1.1 Plans d'expériences optimaux et Mesures dans les réseaux

Les fournisseurs d'accès Internet souhaitent avoir une bonne connaissance du trafic traversant leur réseau, pour de nombreuses opérations contribuant à la bonne gestion du trafic et à la maintenance du réseau. Une partie essentielle de l'information dont ils ont besoin pour ces opérations est la *matrice de trafic*, qui indique les volumes de trafic pour chaque paire origine-destination du réseau pendant un laps de temps donné, c'est à dire le nombre d'octets ayant transité depuis chaque nœud d'entrée vers chaque nœud de sortie pendant la période considérée. L'importance des opérations d'ingénierie du trafic reposant sur la donnée de cette *matrice de trafic* ne cesse d'augmenter, puisque le trafic à traiter augmente et se diversifie, devenant plus complexe d'année en année. Mais en pratique, il est très difficile d'obtenir des estimations précises des demandes de trafic en origine-destination. Contrairement à ce que l'intuition peut laisser croire, les mesures sur les réseaux sont : (i) souvent indisponibles au niveau de certains routeurs non instrumentés ; (ii) coûteuses ; (iii) susceptibles d'affecter la qualité de service. Les décisions concernant l'emplacement des mesures à prendre, ainsi que leur taux d'échantillonage constituent donc un enjeu crucial.

Nous abordons le problème de l'optimisation des mesures dans les réseaux par une approche fondée sur la théorie des *plans d'expériences optimaux*[1]. Cette théorie étudie comment allouer l'effort expérimental à un ensemble d'expériences disponibles, dans le but de maximiser la qualité de l'estimation d'un *paramètre inconnu*. Si l'on considère chaque localisation possible du logiciel de mesure comme une *expérience*, et la matrice de trafic comme le *paramètre inconnu*, on obtient une formulation de type *plans d'expériences* de notre problème de télécommunications. Cependant, les algorithmes classiques pour la conception optimale d'expériences se révèlent inefficaces sur les grands réseaux, principalement parce que de très grandes matrices entrent en jeu.

Cette observation a été notre motivation principale pour rechercher des algorithmes qui passent à l'échelle en conception d'expériences optimales. Nous avons développé une

1. ou *conception d'expériences optimales*

approche reposant sur la *Optimisation Conique du Second Ordre* (SOCP), une classe de problèmes d'optimisation généralisant la Programmation Linéaire (LP), et qui peuvent être résolus par des méthodes de points intérieurs en un temps bien plus court que les Problèmes d'optimisation Semi-Définie (SDP) de la même taille. Cette approche se révèle particulièrement efficace pour les problèmes où l'on cherche à estimer un petit nombre de fonctions linéaires des paramètres inconnus.

En fait, notre approche ne s'applique pas directement au le problème de télécommunications initial. Cela vient du fait que l'opérateur cherche généralement a estimer l'intégralité de la matrice de trafic (tandis que notre approche par SOCP est la mieux adaptée pour l'estimation d'une combinaison linéaire des volumes de trafic). Pour résoudre ce problème, nous avons introduit une méthode pour l'estimation de tous les paramètres du modèle, qui repose repose sur le calcul de plusieurs plans $\boldsymbol{c}-$optimaux.

Un autre problème est lié aux aspects combinatoires de l'application industrielle : si un opérateur souhaite instrumenter un certain nombre de nœuds du réseau afin qu'ils supportent un logiciel de mesure, la formulation naturelle pour choisir quel nœud du réseau équiper en priorité est un *plan d'expériences optimal en nombre entiers*. Ce problème est principalement traité par des approches heuristiques dans la littérature. Ceci a motivé notre travail sur la sous-modularité des critères d'information pour les plans optimaux, et a conduit à des résultats d'approximabilité en temps polynomial de certains problèmes NP-difficiles.

1.2 Organisation et contributions de ce manuscrit

Cette thèse est organisée en deux parties. La première partie est consacrée à des résultats théoriques et algorithmiques en conception optimale d'expériences, qui reposent sur des outils de programmation mathématique et d'optimisation sous-modulaire. Ces résultats ont émergé d'un problème industriel concernant les réseaux de télécommunication, dont l'étude fera l'objet de la seconde partie de ce manuscrit. Nous détaillons ci-dessous le contenu de cette thèse, chapitre par chapitre. Nous dresserons ensuite une liste des contributions de ce manuscrit.

1.2.1 Résumé détaillé

Première Partie : Plans d'expériences optimaux

Dans la première partie, nous présentons des résultats théoriques pour le calcul de plans d'expériences optimaux. Nous nous focaliserons sur les modèles de régression linéaire où le nombre d'expériences disponibles est fini, et nous mettrons l'accent sur le cadre *multiréponses*. Ce dernier correspond à la situation dans laquelle une seule et même expérience peut fournir plusieurs observations simultanées du paramètre inconnu. Les deux premiers chapitres de cette partie reprennent essentiellement l'état de l'art sur la théorie des plans d'expériences optimaux. Les chapitres suivants (4–7) contiennent de nouveaux résultats.

Chapitre 2 : Une introduction à la théorie des plans d'expériences optimaux La théorie des *plans d'expériences optimaux* est une branche importante des statistiques, à l'interface avec l'optimisation, qui a de nombreux champs d'applications. Son but est de trouver les valeurs qu'un expérimentateur doit donner aux *variables de contrôle* des expériences à sa disposition, *avant de les réaliser*. Ces variables de contrôle peuvent prendre différentes formes (nombre de fois qu'on va réaliser une expérience, taux d'échantillonage d'un appareil de mesure, temps pendant lequel on enregistre des résultats, etc.), et affectent les données mesurées. L'estimation que l'expérimentateur fait des quantités qu'il souhaite mesurer va donc dépendre de ces variables.

Dans ce chapitre, nous passons en revue un certain nombre de résultats classiques en conception optimale d'expériences. Nous nous focalisons sur les modèles de régression linéaires, où l'espérance de chaque quantité mesurée est une combinaison linéaire des paramètres inconnus. Nous nous plaçons en outre dans le cadre où une seule et même expérience peut fournir plusieurs mesures simultanées : ce cadre *multiréponses* intervient naturellement dans l'étude du problème de télécommunications traité en Partie II. Nous nous concentrons sur la théorie des *plans approchés*, où la variable de conception est un vecteur $\boldsymbol{w}$ de somme 1, qui indique le pourcentage d'effort expérimental alloué à chaque expérience. Dans le cas où l'ensemble des expériences disponibles $\mathcal{X}$ (l'espace de régression) est infini, l'expérimentateur doit également choisir le sous-ensemble des expériences $\boldsymbol{x_1}, \ldots, \boldsymbol{x_s} \in \mathcal{X}$ à réaliser.

Ce chapitre débute par une rétrospection historique de la théorie des plans d'expériences optimaux, avec une présentation succincte des contributions d'Elfving, Kiefer, Fedorov et Pukelsheim (entre autres). Nous introduirons ensuite la notation standard, et nous montrerons que le théorème de Gauss-Markov donne une borne inférieure pour la matrice de covariance de tout estimateur linéaire sans biais du vecteur des paramètres inconnus. De plus, cette borne est atteinte par l'estimateur des moindres carrés. Ceci conduit à la définition de la *matrice d'information* d'un plan d'expériences (l'inverse de la meilleure matrice de covariance possible), et à la formulation standard des problèmes de conception optimale d'expériences (maximisation d'une fonction scalaire de la matrice d'information). Nous passerons ensuite en revue les critères d'information les plus utilisés dans la littérature, et qui permettent de définir les concepts de $\boldsymbol{c}$, A, E, D, T, Φ_p-optimalité, et de $S-$optimalité robuste.

La dernière partie de ce chapitre rappelle quelques résultats fondamentaux en conception optimale d'expériences :

- Le théorème d'Elfving, qui donne une caractérisation géométrique simple de la $\boldsymbol{c}-$optimalité.
- Le théorème de Kiefer-Wolfowitz (1960), qui montre que le problème de conception $D-$optimale est équivalent à un problème dual (appelé $G-$optimal), et donne une condition nécessaire et suffisante d'optimalité, facile à vérifier en pratique.
- Le théorème d'équivalence général, découvert par Kiefer (1974) et étendu par Pukelsheim (1980), qui généralise le résultat précédent à une large classe de critères d'information.

- Plusieurs conséquences du théorème d'équivalence général, comme des bornes pour les poids en conception $D-$optimale où une formule explicite du plan $A-$optimal quand les vecteurs de régressions forment une famille libre.

Chapitre 3 : Algorithmes classiques pour le calcul de plans optimaux De nombreux algorithmes ont été proposés pour le calcul de plans d'expériences optimaux. Nous en présentons certains dans ce chapitre. Nous restreignons notre étude au cas où le nombre d'expériences est fini (où lorsque les expériences optimales sont données), de sorte que seul le vecteur de poids $\boldsymbol{w}$ entre en jeu dans le problème d'optimisation, ce qui rend le problème convexe. Ce cadre correspond à celui du problème telecoms étudié dans la seconde partie, puisque le logiciel de mesures ne peut être activé que sur un ensemble (fini) de points du réseau.

Le premier algorithme que nous étudions est celui de Fedorov and Wynn pour le calcul de plans $D-$optimaux. Cet algorithme s'inspire du théorème de Kiefer-Wolfowitz : le principe consiste à partir d'un plan d'expériences arbitraire, puis de se déplacer à chaque itération dans une direction donnée par l'évaluation du critère de $G-$optimalité. Le théorème de Kiefer-Wolfowitz garantit qu'il s'agit d'une direction de descente. En fait, cet algorithme appartient à la classe des méthodes de *descentes faisables*. Nous présentons l'extension de cet algorithme a d'autres critères d'information et quelques résultats de convergence.

Nous présentons ensuite la classe des algorithmes multiplicatifs introduits par Titterington. Dans ces algorithmes, l'ensemble des poids du plan d'expériences est mis à jour à chaque itération, en les multipliant chacun par un facteur proportionnel au gradient du critère d'information qu'on maximise. Nous présentons l'algorithme original de Titterington et certaines de ses variantes, ainsi que des résultats récents concernant la convergence de ces méthodes, obtenus par Dette, Pepelyshev et Zhigljavsky (2008) et Yu (2010).

Enfin, nous passons en revue les formulations basées sur l'optimisation semi-définie (SDP) pour les problèmes de plans d'expériences optimaux. Les méthodes de points intérieurs pour résoudre ces problèmes d'optimisation semi-définie sont en général plus lentes que les algorithmes multiplicatifs, mais l'approche SDP offre une grande flexibilité. En particulier, l'utilisateur peut ajouter « sans effort » des contraintes sur les plans d'expériences. Nous donnerons plusieurs exemples des avantage de l'approche SDP.

Chapitre 4 : Un théorème de réduction du rang en Optimisation Semi-définie Ce chapitre contient les résultats de [Sag09a], et présente un intérêt indépendemment du reste de ce manuscrit. Le résultat principal affirme qu'une classe de problèmes d'optimisation semi-définie –qui comprend la plupart des SDP étudiés au Chapitre 3– admet des solutions de petit rang. En fait, l'intuition de ce résultat nous a été donnée par l'extension du théorème d'Elfving au cadre multiréponses (Chapitre 5). Nous avons néanmoins choisi d'insérer ce chapitre à cet endroit du manuscrit, car le théorème principal va s'avérer utile dans plusieurs preuves du Chapitre 5, et mettre en lumière notre approche basée sur l'optimisation conique du second ordre.

La classe des problèmes considérés est celle des *programmes de packing semi-définis*, qui sont les analogues SDP des problèmes de *packing* classiques en programmation linéaire. Notre résultat montre que si la matrice qui définit la fonction objectif du SDP est de rang r, alors le programme de packing semi-défini a une solution dont le rang est inférieur à r. Une conséquence intéressante est le cas dans lequel $r = 1$, car la variable optimale X du SDP peut alors se factoriser sous la forme $X = \boldsymbol{x}\boldsymbol{x}^T$, et nous montrons que trouver $\boldsymbol{x}$ revient à résoudre un problème d'optimisation conique du second ordre (qui est plus simple que le SDP initial).

La preuve de notre résultat peut en fait s'étendre à une classe de problèmes plus large, dans laquelle toutes les contraintes ne sont pas de type *packing*. Nous présentons également cette version étendue de notre résultat.

Chapitre 5 : L'approche par Optimisation Conique du Second Ordre Ce chapitre reprend les résultats de [Sag09b]. Nous montrons que de nombreux problèmes en conception optimale d'expériences peuvent être formulés grâce à l'optimisation coniques du second ordre (SOCP). Contrairement aux approches SDP vues au Chapitre 3, l'approche par SOCP reste efficace pour de très grandes instances, et combine ainsi les avantages de flexibilité des SDP avec la performance des algorithmes multiplicatifs.

Nous commençons par donner une extension du théorème d'Elfving. Ce résultat caractérise géométriquement les plans $\boldsymbol{c}-$optimaux dans le cadre *multiréponses* : les poids optimaux peuvent être lus à l'intersection d'une droite vectorielle et de la bordure de l'enveloppe convexe d'un ensemble d'ellipsoïdes. Nous montrons ensuite que tout problème de plan $A-$optimal se ramène à un problème de plan $\boldsymbol{c}-$optimal, pour des matrices d'observations augmentées. Notre résultat fournit donc une caractérisation géométrique des plans $A-$optimaux.

Nous mentionons toutefois qu'un résultat équivalent a été formulé de façon indépendante par Dette et Holland-Letz en 2009, dans un cadre différent. Dette et Holland-Letz ont considéré un modèle hétéroscedastique (c'est à dire un modèle où la moyenne *et la variance* des observations sont des fonctions du paramètre inconnu). Ce modèle peut se ramener à considérer des matrices d'observations de rang $k \geq 2$, de façon similaire au modèle des expériences *multiréponses*. Nous proposons une preuve et une analyse des conséquences de ce résultat différentes de celles de Dette et Holland-Letz.

Un corollaire de cette extension du théorème d'Elfving est une formulation SOCP du problème de plan $\boldsymbol{c}-$ (ou $A-$) optimal pour des expériences multiréponses. Nous donnons une seconde preuve de cette réduction basée sur le théorème du Chapitre 4 : Le SDP pour la $\boldsymbol{c}-$optimalité a une solution de rang 1, et se ramène à un SOCP. De façon plus générale, nous verrons que les problèmes de conception $A-$optimale où le plan d'expériences est sujet à plusieurs contraintes linéaires admettent une formulation SOCP. Là encore, nous donnons deux preuves de ce résultat, l'une s'appuyant sur un argument de statistiques et l'autre sur notre théorème de réduction du rang.

Nous nous intéressons ensuite à d'autres critères d'optimalité. Nous montrons que le problème de plan $T-$optimal pour un sous-système des paramètres inconnus se ramène lui aussi à un SOCP. Enfin, nous considérons le critère robuste de $S-$optimalité introduit par Läuter ; le problème de plan optimal correspondant se ramène à la minimisation d'une moyenne géométrique sous des contraintes de type SOCP. En suivant une approche similaire à celle de Dette (1993), nous obtenons alors une formulation SOCP pour le problème de conception $D-$optimale. De plus, nous montrons que les conditions d'optimalité de notre programme géométrique généralisent un théorème de Dette (1993) au cadre multiréponses.

Chapitre 6 : Comparaison numériques des algorithmes Nous évaluons dans ce chapitre les bénéfices de notre approche par SOCP pour le calcul des plans d'expériences optimaux. Notre approche se révèle très efficace pour plusieurs critères d'optimalité, surtout lorsque le nombre r de fonctions linéaires des paramètres que l'on cherche à estimer est petit (en particulier pour le problème de plan $\boldsymbol{c}-$optimal).

Nous comparons notre approche avec les algorithmes classiques présentés au Chapitre 3, à savoir les algorithmes d'échange de type Wynn–Fedorov, les algorithmes multiplicatifs *à la* Titterington, et l'approche par optimisation semi-définie.

Plusieurs types d'instances sont considérées. Dans un premier temps, nous étudions des instances aléatoires, dans le but d'évaluer dans quelle mesure les différents paramètres (nombre d'expériences, nombre d'inconnues, critère maximisé, nombre de fonctions linéaires que l'on cherche à estimer,...) affectent le temps de calcul. Nous nous intéressons ensuite à des problèmes de régressions polynomiales, qui ont été très étudiés dans la littérature sur les plans d'expériences. Nous présentons enfin quelques résultats numériques sur des instances provenant de l'application aux réseaux qui fait l'objet de la seconde partie de ce manuscrit.

Chapitre 7 : Problèmes combinatoires en conception optimale d'expériences Ce chapitre présente les résultats de [Sag10]. Certains résultats avaient également été annoncés dans [BGS08]. Nous nous intéressons aux aspects combinatoires dans les problèmes de plans d'expériences optimaux. Dans de nombreuses applications, les variables contrôlant les plans d'expériences sont discrètes, voire binaires. Ce chapitre fournit des résultat d'approximabilité en temps polynomial pour le problème de conception optimale d'expériences en nombres entiers, qui est NP-difficile.

En particulier, nous établissons une inégalité matricielle qui montre que la fonction objectif du problème d'optimisation considéré est *sous-modulaire*. Nous en déduisons que l'approche gloutonne –qui a souvent été utilisée pour ce problème– fournit toujours un plan d'expériences qui approche l'optimum par un facteur d'au moins $1 - 1/e \approx 62\%$. Notre résultat d'approximabilité peut également s'étendre au cas où les expériences n'ont pas toutes le même coût.

Nous étudions ensuite les algorithmes consistant à arrondir la solution du problème relâché continu, une approche qui a été appliquée par de nombreux auteurs. Lorsque l'on

souhaite choisir un sous-ensemble de n parmi s expériences, nous montrons que le plan $D-$optimal peut être arrondi aléatoirement, de façon à obtenir un plan d'expérience *entier*, pour lequel la dimension du sous-espace observable approche l'optimum par un facteur $\frac{n}{s}$ avec une grande probabilité. Si ce résultat peut sembler plus faible que le résultat d'approximation gloutonne précédent, nous montrons néanmoins que le facteur $\frac{n}{s}$ est (presque) optimal, car il y a des instances pour lesquelles le ratio d'approximabilité moyen est de $\frac{n}{s-1}$

Seconde Partie : Contrôle optimal des grands réseaux

Dans la seconde partie de ce manuscrit (page 145), nous étudions une application de la théorie des plans d'expériences optimaux pour le contrôle optimal des grands réseaux *backbone*. Les fournisseurs d'accès à Internet souhaitent surveiller le trafic sur leur réseau pour plusieurs raisons. Dans cette thèse, nous nous concentrons sur l'une d'entre elles uniquement : le problème de l'estimation la plus précise possible de la matrice de trafic. Cette matrice donne le volume de trafic pour chaque paire Origine-Destination du réseau, et est nécessaire pour de nombreuses opérations contribuant à la bonne gestion du trafic et à la maintenance du réseau. Nous pensons que notre approche (optimisation des mesures pour l'estimation de la matrice de trafic) est bien fondée car elle indique comment choisir les mesures afin de capturer le plus d'information possible sur le trafic dans le réseau.

Les deux premiers chapitres de la seconde partie présentent l'état de l'art sur l'estimation des matrices de trafic dans les réseaux IP (Chapitre 8), avec un accent particulier sur les approches basées sur la théorie de l'information et les projections entropiques, ainsi que leur rapport historique avec les problème de *matrix balancing* (Chapitre 9). Le chapitre 10 contient les principaux résultats de cette partie, et des perspectives sont présentés au Chapitre 11.

Chapitre 8 : Estimation des matrices de trafic : État de l'art L'estimation des matrices de trafic dans les réseaux a fait l'objet de recherches intensives pendant la dernière décennie, de la part des opérateurs Internet et de la communauté académique travaillant sur les réseaux. Dans ce chapitre, nous passons en revue les différentes méthodes qui ont été proposées pour faire cette estimation. On peut principalement les séparer en deux catégories : les méthodes qui n'utilisent que les mesures sur les liens, et celles qui se fondent sur des mesures directes des volumes de trafic en origine-destination enregistrées par un logiciel de contrôle.

L'inférence de la matrice de trafic à partir des mesures sur les liens est un problème classique, très pur d'un point de vue mathématique : étant donné un réseau avec son ensemble de liens, et un ensemble de paires origine-destination (OD) qui empruntent ces liens (le chemin utilisé pour chaque OD est supposé connu), le problème est de trouver comment se répartit le volume total de trafic parmi les paires OD, cette répartition devant être cohérente avec les volumes observés sur chaque lien. Ce problème est typiquement sous-déterminé, puisque sur un réseau avec n nœuds, le nombre de liens est de l'ordre de n tandis que le nombre d'inconnues (les volumes de trafic sur chaque OD) est d'ordre n^2.

Pour résoudre ce problème, des méthodes Bayesiennes ou basées sur la théorie de l'information ont été proposées. Dans l'approche Bayesienne, on suppose que la matrice de trafic suit une un loi paramétrique, et on maximise la vraisemblance des mesures sur les liens pour choisir la valeur des paramètres. Cette maximisation peut se faire, par exemple, avec l'algorithme Espérance-Maximisation. L'approche basée sur la théorie de l'information se ramène à résoudre des problèmes de maximisation d'entropie, qui seront étudiés en détail au Chapitre 9.

Les méthodes les plus évoluées se basent sur des mesures directes des volumes de trafic en OD, enregistrées par un logiciel comme Netflow de Cisco Systems. Pour des raisons que nous détaillerons dans ce chapitre, l'utilisation intensive de Netflow n'est cependant pas souhaitable. Là encore, on peut séparer les méthodes d'estimation de la matrice de trafic en deux catégories : il a été proposé d'une part d'utiliser Netflow de façon intensive pendant une certaine période seulement, pour construire un modèle précis des demandes de trafic. Ce modèle est ensuite utilisé pour estimer la matrice de trafic à des temps ultérieurs où Netflow est désactivé. Le modèle doit être recalibré au bout d'un certain temps, car le trafic n'est pas stationnaire. Cette classe de méthodes utilisant Netflow pour la calibration d'un modèle du trafic regroupe, entre autres, la technique du filtre de Kalman, l'analyse en composantes principales, et la méthode des *fanouts*. Leur inconvénient commun est la durée des périodes de recalibration, qui est relativement longue (au moins 24 heures de mesures intensives sont nécessaires). D'autre part, des méthodes récentes utilisent des mesures partielle de Netflow, enregistrées de façon régulières, mais au niveau d'un petit nombre de routeurs seulement. Nous présentons brièvement l'ensemble de ces méthodes et nous les comparons sous plusieurs critères.

Chapitre 9 : Théorie de l'information et projections entropiques Dans l'approche basée sur la théorie de l'information, nous normalisons la matrice de trafic de sorte qu'elle somme à 1. La matrice ainsi obtenue peut s'interpréter comme la distribution de probabilité qu'un paquet choisi au hasard appartienne à telle ou telle paire OD. En suivant le principe de maximisation d'entropie, la distribution de probabilité qui représente le mieux l'état de notre connaissance est, parmi l'ensemble des distributions qui vérifient les équations de mesures sur les liens, celle avec la plus grande entropie. Cette approche justifie le modèle *gravitaire* de la matrice de trafic, qui est la matrice de trafic avec l'entropie maximale lorsque les seules mesures disponibles sur le réseau sont sur les liens externes (liens d'entrées et de sortie) – c'est à dire lorsque le comportement interne du réseau est représenté par une boîte noire.

Le modèle gravitaire peut être utilisé comme une bonne estimation *a priori* de la matrice de trafic. Toujours en suivant la théorie de l'information, une approche naturelle consiste à choisir la distribution des volumes de trafic qui satisfait les équations de mesures, et est la plus difficile à distinguer de l'estimation *a priori*. Cette approche conduit à formuler des problèmes de *projections entropiques* où l'on minimise la divergence de Kullback-Leibler entre les volumes de trafic et l'estimation gravitaire, sous les contraintes imposées par les mesures au niveau des liens.

Nous présentons ensuite quelques résultats sur ce problème d'optimisation, dont une partie a été obtenue au cours d'un stage de recherche précédent la présente thèse. Nous montrons que les points stationnaires sont en correspondance avec les racines d'un système d'équations polynomiales linéaires en chaque variable. Nous donnons des conditions simples qui garantissent l'existence et l'unicité de la solution de ce système. En particulier, nous analysons la similarité entre l'algorithme classique "Iterative proportional fitting" (IPF) –qui a souvent été utilisé pour le problème d'inférence de la matrice de trafic– et les algorithmes classiques de *matrix balancing*. Nous montrons que la généralisation directe des algorithmes de *matrix balancing* aux projections entropiques dans les réseaux ne fonctionne que si toutes les paires OD sont de longueurs inférieures ou égales à 2. Dans l'algorithme IPF, les variables sont mises à jour une à une, de façon cyclique (au lieu d'être modifiée simultanément comme dans les problèmes de *balancing*). Cette différence fait de l'IPF un algorithme de projections cycliques, et on sait en conséquence qu'il a un taux de convergence linéaire.

Chapitre 10 : Optimisation des mesures Netflow Ce chapitre présente plus en détails les résultats de [SBG10, SGB10]. Nous montrons que le problème consistant à trouver les localisations optimales de Netflow, ainsi que celui de choisir les meilleurs taux d'échantillonnages, peuvent se formuler sous la forme de problèmes standards de plans d'expériences optimaux. Le problème principal est la taille des matrices impliquées dans ce problème, qui sont de taille $n^2 \times n^2$ pour un réseau avec n nœuds. Quand $n \geq 17$, les approches semi-définies deviennent alors inefficaces.

Nous proposons une nouvelle procédure, que nous avons appelée "plans $\boldsymbol{c}-$optimaux successifs" (PCOS), dans lequel un plan d'expérience est construit en prenant la moyenne de plusieurs plans $\boldsymbol{c}-$optimaux. Cette approche a l'avantage de *très bien passer à l'échelle*. Il est à souligner que des éléments heuristiques laissent penser que lorsque les vecteurs $\boldsymbol{c}$ sont tirés selon une loi Gaussienne, la limite théorique du plan construit par l'approche PCOS est proche du plan $A-$optimal. Nous montrons des exemples où cette affirmation est vérifiée en pratique.

De nombreux réseaux ne sont pas (ou seulement partiellement) instrumentés avec Netflow. Lorsqu'un opérateur décide d'équiper un nombre additionnel de routeurs avec Netflow, le problème est de choisir quels routeurs instrumenter en priorité. Nous comparons notre approche (PCOS) avec l'algorithme glouton pour le problème de déploiement de Netflow. Toutes nos expériences sont basées sur des données réelles provenant des réseaux *Abilene* et *Opentransit* (ce dernier est le backbone international de France Telecom).

Nous adaptons ensuite notre approche pour prendre en compte les mesures prises à des instants antérieurs (dans un contexte dynamique, l'opérateur peut ne pas avoir intérêt à activer Netflow avec des hauts taux d'échantillonage sur la même interface pendant des périodes successives ; si un point du réseau est bien mesuré à l'instant t, il semble intuitif de concentrer l'effort de mesure à un autre endroit du réseau à $t+1$). Pour ce faire, nous utilisons un article récent de Singhal and Michailidis. Ces auteurs ont formulé un problème de plan optimal dans lequel la matrice d'information comprend un terme supplémentaire

pour les erreurs des mesures passées qui est mis à jour à chaque pas de temps grâce à un filtre de Kalman. En fait, nous montrons par un exemple sur Abilene qu'en raison de la grande variabilité du trafic, il est parfois préférable d'ignorer l'effet des mesures passées.

Finalement, nous évaluons notre approche pour le problème d'échantillonage optimal avec Netflow, pour le cas de contraintes *par routeur*. Étant donné un nombre maximal de paquets que Netflow peut analyser au niveau de chaque routeur, le but est de trouver la répartition optimale des mesures au niveau de chaque routeur, c'est à dire régler au mieux les taux d'échantillonage sur chaque interface tout en maintenant le nombre de paquets échantillonnés sous le seuil autorisé. Nous étudions par notre approche PCOS une instance de ce problème sur le réseau Opentransit, qui comprend 13456 paires OD, 116 routeurs et 436 interfaces. Nous ne connaissons pas d'autres approches qui pourraient traiter des instances de cette taille.

Chapitre 11 : Perspetives pour la modélisation spatio-temporelle des matrices de trafic Nous présentons dans ce chapitre quelques perspectives pour l'estimation des matrices de trafic. Il s'agit d'un travail préliminaire, basé sur la théorie des matrices aléatoires et des décompositions de petit rang des tenseurs.

Quand on la considère au cours du temps, la matrice de trafic est en fait un objet tridimensionnel (origines x destinations x temps), qui a presque toujours été traité comme un objet à deux dimensions par les auteurs de la communauté réseaux. Pour se ramener à des matrices, les matrices origine-destination sont vectorisées sous la forme d'un vecteur colonne à chaque pas de temps. Cependant, cette vectorisation fait perdre une précieuse information sur les corrélations qui existent entre les origines et les destinations.

Nous avons étudié la distribution empirique des valeurs singulières des matrices de trafic OD, à partir des données réelles dont nous disposons sur Abilene et Opentransit. Il est intéressant de remarquer que mise à part quelques grandes valeurs singulières, la distribution du bas du spectre correspond très bien à la distribution théorique que devrait avoir le spectre d'une matrice aléatoire, dite de *Wishart*. Cette remarque laisse penser que chaque matrice origine-destination peut se décomposer comme la somme d'une matrice de petit rang (qui supporte la partie déterministe du signal), plus une matrice de bruit aléatoire, dont la distribution est reliée à la loi de Wishart. Cette étude préliminaire n'est pas encore une méthode pour filtrer le bruit et estimer les matrices de trafic à partir de mesures incomplètes. En revanche, il nous semble essentiel de modéliser la structure de petit rang des matrices de trafic Origine-Destination. C'est chose faite dans la dernière section de ce chapitre, consacrée à l'étude de décompositions de petit rang du tenseur de trafic tridimensionnel.

Si les approximations de petit rang de matrices sont des problèmes parfaitement compris de nos jours (grâce aux troncations de la décomposition en valeur singulières), les approximations de petit rang des tenseurs font en revanche l'objet de recherches actives. Nous passons en revue quelques résultats et algorithmes sur les décompositions de tenseurs, et nous montrons le potentiel de ces méthodes en analysant les décompositions de tenseur

de trafic avec des données réelles (Abilene et Opentransit). Finalement, nous présentons l'esquisse d'une méthode –basée sur les décompositions tensorielles– pour l'estimation *en ligne* des matrices de trafic à partir de mesures incomplètes. Nous montrons par un exemple sur Opentransit que notre méthode conduit à une amélioration par rapport à la méthode classique *tomogravitaire*.

1.2.2 Contributions de cette thèse

Nous listons ci-dessous les contributions principales de cette thèse :

- Théorème 4.1.2, et son extension Théorème 4.2.2. Tout problème de la classe des *programmes de packing semi-définis* où la matrice dans la fonction objectif est de rang r a une solution de rang inférieur ou égal à r. Nous discutons les extensions et conséquences de ce résultat. Ce théorème sera utilisé plusieurs fois au Chapitre 5.
- Théorème 5.1.1 : Extension du théorème d'Elfving au cadre multiréponses (Nous avons présenté ce résultat à la conférence [SBG09]. Il a été découvert de façon indépendante par Dette et Holland-Letz [DHL09]).
- Théorème 5.2.1 : Formulation SOCP du problème de plan $\boldsymbol{c}-$optimal. Nous donnons une interprétation géométrique de ce résultat.
- Extension du résultat précédent au critère de $A-$optimalité (Théorème 5.2.2), et au cas où le plan d'expériences est soumis à plusieurs contraintes linéaires (Théorème 5.2.3).
- Théorème 5.2.5 : Formulation SOCP du problème de plan $T-$optimal pour un sous-système de paramètres $K^T\boldsymbol{\theta}$.
- Théorème 5.3.1 : Formulation sous forme d'un programme géométrique du problème robuste de $S_\beta-$optimalité. Les conditions d'optimalité de ce problème généralisent un résultat de Dette [Det93] au cadre multiréponses (Théorème 5.3.2).
- Un corollaire du résultat précédent est un SOCP pour le problème de plan $D-$optimal (cf. Equation 5.25).
- Tests numériques et comparaisons avec d'autres algorithmes (Chapitre 6), montrant l'efficacité de l'approche par SOCP lorsque le nombre r de fonctions linéaires des paramètres à estimer est petit (en particulier pour les plans $\boldsymbol{c}-$optimaux où $r = 1$).
- Théorème 7.2.1 : Réduction du problème combinatoire de *plans d'expériences de rang maximal* à MAXCOVERAGE. En conséquence, si l'on admet $P \neq NP$, il n'existe pas d'algorithme polynomial qui approche le *plan de rang maximal* par un facteur plus grand que $1 - e^{-1}$.
- Proposition 7.2.4 : Si f' est operateur antitone sur $\mathbb{R}_+^*$, alors pour tout triplet $(X, Y, Z) \in \mathbb{S}_m^+$

$$\text{trace } f(X + Y + Z) + \text{trace } f(Z) \leq \text{trace } f(X + Z) + \text{trace } f(Y + Z).$$

- Corollaire 7.2.6 : Le critère Φ_p de Kiefer (vu comme une fonction ensembliste) est *sous-modulaire croissant* pour $p \in [0, 1]$.

- Théorème 7.2.7 : En conséquence, l'algorithme glouton retourne toujours une solution approchant par un facteur d'au moins $1 - e^{-1}$ l'optimum du problème de plan Φ_p-optimal (pour $p \in [0, 1]$). Des extensions possibles de ce théorème sont présentées.
- Proposition 7.3.4 (cf. également Théorème 2.4.7) : Généralisation des bornes supérieures pour les poids $D-$optimaux au cadre multiréponses (découvert indépendemment par Harman et Trnovská [HT09] pour le cas de l'estimation du vecteur complet des paramètres $\boldsymbol{\theta}$, i.e. quand $K = \boldsymbol{I}$).
- Théorème 7.3.7 : Si l'on doit choisir n expériences parmi s, nous donnons deux algorithmes d'arrondi randomisé qui retournent une solution approchant l'optimum du problème de *plan de rang maximal* par un facteur n/s (en moyenne).
- Nous montrons des instances pour lesquelles le ratio d'approximation des algorithmes randomisés précédents est $n/(s-1)$ (cf. Remarque 7.3.2).
- Proposition 9.5.7 : Pour le problème de projection entropique avec contraintes linéaires sur un réseau, l'algorithme de point fixe naturel est contractant si et seulement si toutes les paires OD sont de longueurs inférieures ou égales à 2. (Résultat obtenu pendant un stage antérieur à cette thèse.)
- Formulation de type plan d'expériences pour le problème du déploiement optimal de Netflow, et le problème de l'échantillonnage optimal de Netflow (cf. Section 10.2).
- Proposition d'une nouvelle méthode (baptisée *Plans $\boldsymbol{c}-$Optimaux Successifs*, PCOS) basée sur le calcul de plusieurs plans $\boldsymbol{c}-$optimaux pour traiter les problèmes de grande taille en conception d'expériences (cf. Section 10.4.1). Ebauche d'une justification heuristique de notre approche (Sections 10.4.2 et 10.4.3).
- Validation de notre approche par des tests utilisant des données réelles (cf. Section 10.5).
- Mise en évidence de la structure de petit rang des matrices de trafic origine-destination. Proposition d'un modèle signal + bruit, et analyse préliminaire du *bruit* par des outils de la théorie des matrices aléatoires (cf. Section 11.1).
- Mise en évidence de la structure de petit rang des *tenseurs de trafic* tridimensionels (origines $\times$ destinations $\times$ temps). Esquisse d'une méthode reposant sur les tenseurs pour estimer les matrices de trafic *en ligne* (cf. Section 11.2.3).

Introduction (in English)

This chapter briefly presents our motivation and the scientific path which has led to this thesis. At the end of this chapter, we draw a detailed outline and list the contributions of this thesis.

1.3 Optimal design of experiments and Network measurements

Internet Service Providers (ISP) wish to have a good knowledge about the traffic which transit through their networks, for many traffic engineering and network planning tasks. An essential part of the required information is the *traffic matrix*, which contains the volumes of traffic for each origin-destination pair of the network during a given period of time, i.e. the number of bytes that has travelled from any entry node to any exit node. The importance of the networking operations relying on the traffic matrix is increasing as the traffic grows in volume and becomes more complex, but in practice, obtaining accurate estimations of the demands of traffic is a challenging issue. Contrarily to what intuition may suggest, network measurements are: (i) often not available everywhere; (ii) expensive; (iii) likely to affect the quality of service. It is thus a crucial issue to decide where network measurements should be performed, as well as their sampling rates.

We approach the problem of optimizing the network measurements by using the theory of *optimal experimental designs*[2]. This theory studies indeed how to allocate the experimental effort to a set of available experiments, in order to maximize the quality of estimation of an *unknown parameter*. Thinking of each potential location of the measuring software as an *experiment*, and the traffic matrix as the *unknown parameter*, one obtains a nice *optimal experimental design* formulation of our telecommunications problem. However, the classic optimal experimental design algorithms are intractable on large scale networks, because very large matrices are involved.

This observation motivated us to search for scalable algorithms in optimal experimental design. We developed an approach relying on *Second Order Cone Programming* (SOCP), a class of mathematical optimization problems which generalizes Linear Programs (LP), and which can be solved by interior-point methods in a much shorter time than Semidefinite

2. or theory of optimal experiments

Programs (SDP) of the same size. This approach turns out to be very efficient for problems in which a small number of linear functions of the unknown parameter must be inferred.

In fact, our approach can not be directly applied to the initial telecommunications problem. The reason is that the ISP usually wishes to estimate the whole traffic matrix (while our SOCP approach is best-suited for the estimation of a linear combination of the volumes of traffic). To overcome this problem, we have proposed a new method which rely on the computation of several $\boldsymbol{c}-$optimal designs, and can be efficiently implemented by solving a sequence of SOCP.

Another issue arising from the industrial problem is the combinatorial aspect: when an ISP wishes to upgrade a set of routers of the network, so that they can support the measuring device, the natural formulation is an *integer optimal design problem*. This problem is mainly handled by heuristic approaches in the literature, which motivated our work on the submodularity of the experimental design information criteria. This approach led to polynomial-time approximability bounds for some NP-hard optimization problems.

1.4 Organization and contributions of this manuscript

This thesis is organized in two different parts. The first part is devoted to theoretical and algorithmic results in optimal experimental design, which rely on mathematical programming and submodular optimization. These results have emerged from an industrial problem in telecommunication networks, which we study in the second part of this manuscript. We detail below the content of each chapter. Then, we shall list the contributions of this thesis.

1.4.1 Detailed outline

Part I: Optimal Design of Experiments

In a first part, we present theoretical results for the numerical computation of optimal experimental designs. The focus is on linear regression models, when the number of available experiments is finite, with a special interest for the situation in which one single experiment can produce several observations at the same time (*multiresponse* framework). The first two chapters of this part essentially recall the necessary background on the theory of optimal experimental designs. The following chapters (4–7) contain new results.

Chapter 2: An introduction to the theory of Optimal Experiments The theory of *optimal experimental designs* is an important branch of statistics at the interface with Optimization, which has a very wide spectrum of applications. It aims at finding the optimal value that the experimenter should give to the control variables of the experiments at his disposal, *before to perform them*. These control variables (number of times that we perform a measurement,

sampling rate of a device, time at which the measurement will be recorded, etc.) affect the measured data, and so the inference of the quantities of interest depends on those variables.

In this chapter, we review classic results of the theory of optimal experimental design. We focus on the linear regression models, in which the expected value of the measurements is linear with respect to the unknown parameters. In addition, a single experiment is allowed to produce a multidimensional observation: this is the natural setting for the optimal monitoring problem which will be studied in Part II. We concentrate our attention to *approximate designs*, that is, the design variable is a continuous vector $\boldsymbol{w}$ summing to 1 ($\sum_{i=1}^{s} w_i = 1$), which indicates the allocation of the experimental effort to the available experiments. If in addition the set of potential experiments $\mathcal{X}$ (the *regression region*) is infinite, the experimenter should also find the optimal measurement points $\boldsymbol{x_1}, \ldots, \boldsymbol{x_s} \in \mathcal{X}$ where to perform the experiments.

This chapter starts with a historical review of the theory of optimal experimental design, with a brief presentation of the contributions of Elfving, Kiefer, Fedorov and Pukelsheim (among others). We next introduce the standard notation, and we shall see that the Gauss-Markov theorem gives a *lower bound* on the covariance matrix for an unbiased estimator of the parameters, which is attained for the least-square estimator. This yields the definition of the *information matrix* of a design (as the inverse of this best variance), and the general formulation of the optimal design problem, i.e. the *maximization of a scalar function of the information matrix*. We next review the popular information criteria from the optimal experimental design literature, which define the concepts of $\boldsymbol{c}$, A, E, D, T, Φ_p, and robust $S-$optimality.

The last part of this chapter is devoted to a review of some fundamental results in optimal experimental design:

- The Elfving theorem (1952), which gives a simple geometric characterization of $\boldsymbol{c}-$optimality.
- The Kiefer-Wolfowitz theorem (1960), which shows that the $D-$optimal problem is equivalent to a dual problem (called $G-$optimal) and gives optimality conditions that one may easily check.
- The general equivalence theorem, discovered by Kiefer (1974) and extended by Pukelsheim (1980), which generalizes the latter result to a large class of information criteria.
- Some consequences of the general equivalence theorem, like bounds on the $D-$optimal weights or a close form formula of the $A-$optimal design on independent regression vectors.

Chapter 3: Classic algorithms for computing optimal designs Many algorithms have been proposed to compute optimal experimental designs. We review some of them in this chapter. We restrict our study to the case in which the number of available experiments is finite (or the optimal measurement points are given). Thus, the optimization is carried over the vector of weights $\boldsymbol{w}$ only, and the optimization problem becomes convex. This is also the setting

of the optimal monitoring problem studied in Part II, where the monitoring devices may be activated at a finite number of given locations.

The first algorithm that we study is the one of Fedorov and Wynn for the computation of $D-$optimal designs, which was inspired by the the Kiefer-Wolfowitz theorem. The idea is to start from an arbitrary design and to move at each step in a direction which is given by the evaluation the $G-$criterion. The Kiefer-Wolfowitz theorem ensures that this is a descent direction. This algorithm is in fact a feasible descent method. We present the extension of this algorithm to a wider class of information functions and discuss convergence issues.

We next review the class of multiplicative algorithms, introduced by Titterington. The principle of this class of algorithms is to update simultaneously all the weights of a design, by multiplying them by a factor which is proportional to the gradient of the objective function. We present the original algorithm of Titterington and some of its variants, as well as recent convergence results from Dette, Pepelyshev and Zhigljavsky (2008) and Yu (2010).

Finally, we review some semidefinite programming (SDP) formulations of optimal experimental design problems. The interior point algorithms for semidefinite programming are usually slower than the multiplicative update algorithms, but they offer a lot of flexibility, and the possibility to add "without effort" new constraints in the problem. We give several examples of the advantages of the SDP approach.

Chapter 4: A Low rank reduction Theorem in Semidefinite Programming This chapter contains the results of [Sag09a], and is of independent interest. The main result is that a class of semidefinite programs – which encompass the semidefinite programs of Chapter 3 – admits solutions of low rank. In fact, we got the intuition of this result from the extension of Elfving's theorem to the multiresponse framework (Chapter 5). We have chosen to insert this chapter at this point of the manuscript though, because our theorem will provide alternative proofs of the results of Chapter 5, shedding more light on our Second order cone programming approach.

The class of semidefinite programs considered are *semidefinite packing programs*, which are the SDP analogs to the packing problems in linear programming. Our main result states that if the matrix defining the objective function of this SDP has rank r, then the semidefinite packing program has a solution that is of rank at most r. An interesting corollary is the case in which $r = 1$, because the optimal SDP variable X can be factorized as $\boldsymbol{x}\boldsymbol{x}^T$, and we show that finding $\boldsymbol{x}$ reduces to a Second-Order Cone Program (SOCP), which is computationally more tractable than the initial SDP.

The proof of this result actually carries over a wider class of programs, in which not all variables are subject to *packing* constraints. We next present this extended version of our result.

Chapter 5: The Second Order Cone Programming approach This chapter contains the results of [Sag09b]. We show that several optimal experimental design problems may be

formulated as second order cone programs. In contrast to the SDP approach of Chapter 3, the SOCP approach remains tractable and efficient on very large instances, thus combining the performance of multiplicative update algorithms and the flexibility of semidefinite programs.

We start by giving an extension of the Elfving theorem. The result is a geometric characterization of the $\boldsymbol{c}-$optimal designs for multiresponse experiments: the optimal weights can be read at the intersection of a straight line and the boundary of the convex hull of ellipsoids. We next point out that the $A-$optimal design problem can be formulated as a $\boldsymbol{c}-$optimal design problem with augmented observation matrices, such that our result also yields a geometric characterization of $A-$optimality.

It should be mentioned that an equivalent result was established independently by Dette and Holland-Letz in 2009, but in a different context. Dette and Holland-Letz considered a heteroscedastic model (i.e. an experimental model where both the mean and the variance of the observations depend on the parameter of interest), which led them to study the case in which the observation matrices are of rank $k \geq 2$, just as in the model of *multiresponse experiments*. The proof and the analysis of the consequences of the present result presented in this chapter are different than those of Dette and Holland-Letz.

A consequence of this extended Elfving theorem is that the $\boldsymbol{c}-$ (and $A-$) optimal design of multiresponse experiments can be formulated as a second order cone program. We give an alternative proof of this result, relying on the rank reduction theorem of Chapter 4: the $\boldsymbol{c}-$optimality SDP presented in Chapter 3 has a rank-one solution and so it reduces to a SOCP. More generally, we shall see that the $A-$optimal design problem with multiple linear constraints can be formulated as a SOCP. Again, we give two proofs of this result, one relying on a statistical argument and the other one on our rank reduction theorem.

We next investigate other optimality criteria. We shall see that the $T-$optimal design problem for the estimation of a parameter subsystem can also be formulated as a SOCP. Then, we consider the robust $S-$optimality criterion introduced by Läuter: the corresponding optimal design problem reduces to the maximization of a geometric mean under SOCP constraints. As a consequence, we obtain a SOCP for $D-$optimality, by following the approach of Dette (1993). Moreover, we show that the optimality condition of our geometric program generalizes a theorem of Dette (1993) which geometrically characterizes the $S-$optimality.

Chapter 6: Numerical comparison of the algorithms In this chapter, we evaluate the benefits of our SOCP approach for the computation of optimal experimental designs. We shall see that for several optimization criteria, the second order cone programs presented in Chapter 5 are very efficient when the number r of linear functions of the parameter to estimate is small (in particular for $\boldsymbol{c}-$optimality).

We compare our approach to the algorithms presented in Chapter 3, namely semidefinite programs, Wynn–Fedorov-type exchange algorithms, and Titterington-type multiplicative

algorithms.

We consider several kind of instances. At first, we study random instances of optimal design problems, in order to evaluate to which extent each parameter (number of experiments, number of unknowns, number of linear functions to estimate, design criterion,...) affects the computation time. Then, we consider classic polynomial regression models that have been extensively studied in the experimental design literature. Finally we present some computational results from the network application which will be developed in the second part of this thesis.

Chapter 7: Combinatorial problems arising in optimal design of experiments This chapter contains the results presented in [Sag10]. Some of them were already announced in [BGS08]. We investigate combinatorial aspects of the optimal experimental design problems. In a number of real-world applications, the variables controlling the experimental design are discrete, or binary. This chapter provides some polynomial-time approximability results for the discrete optimal experimental design problem, which is NP-hard.

In particular, we establish a matrix inequality which shows that the objective function is submodular, from which we deduce that the greedy approach, which has often been used for this problem, always gives a design within $1 - 1/e \approx 62\%$ of the optimum. Our result also extends to the budgeted case, in which experiments have different costs.

We next study the design found by rounding the solution of the continuous relaxed problem, an approach which has been applied by several authors: When the goal is to select n out of s experiments, we show that the $D-$optimal design may be rounded to a random subset of n experiments for which the dimension of the observable subspace is within $\frac{n}{s}$ of the optimum with a high probability. This result may look disappointing in the first place, but we show that the $\frac{n}{s}-$factor is (almost) optimal since there are some instances for which the average ratio of approximation is $\frac{n}{s-1}$.

Part II: Optimal monitoring in large Networks

In the second part of this manuscript (page 145), we study an application of the theory of optimal experimental designs to the monitoring of large backbone networks. Internet providers want to monitor their networks for several different objectives, but in this thesis we concentrate on the problem of accurately inferring the traffic matrix only: this matrix gives the volume of traffic for every origin-destination pair of the network, and is needed for many networking applications. We believe that this approach is well funded, because it indicates which part of the network captures the most valuable information about the traffic.

The first two chapters of part II present the background on the traffic matrix estimation in IP networks (Chapter 8), with a particular insight into the information theoretic approaches

relying on entropic projections, and their historic relation with matrix balancing problems (Chapter 9). Chapter 10 contains the main results of this part, and Chapter 11 presents some perspectives.

Chapter 8: Inference of the traffic matrix: a review The estimation of traffic matrices in networks has attracted much interest for the last decade, from both Internet providers and the network research community. In this chapter, we review the different methods that have been proposed for this task; they can principally be classified in two types: those relying on the link counts only, and those which take advantage of direct network measurements provided by a monitoring software.

The inference of the traffic matrix from link counts is a classic problem, very pure on a theoretical point of view: given a network with its set of links, and a set of Origin-Destination (OD) pairs routed on these links (the path for each OD is assumed to be known), the goal is to find the repartition of the total volume of traffic between the different OD pairs, such that this allocation is consistent with the volumes observed on the links. This problem is typically underdetermined, since on a network with n nodes, the number of links is in the order of n, while the number of unknown OD flows is typically of order n^2.

To tackle this issue, Bayesian and information theoretic methods have been proposed. In the Bayesian approach, a parametric law is assumed for the distribution of the flow volumes (i.e. the volumes of traffic on the OD pairs), and we select the parameters of this law so as to maximize the likelihood of the observation on the link counts. Typically, this can be carried out by the Expectation-Maximization algorithm. The information theoretic approach leads to entropic projections, which will be studied with more details in Chapter 9.

Some more evolved methods allow the use of direct measurements, which can be collected by a network monitoring tool, like Netflow from Cisco Systems. For technical reasons which we detail in this chapter, the intensive use of Netflow on the network is not suited. Here again, we can separate the estimation methods in two types, depending on the measuring scheme: on the one hand, some methods require an intensive use of Netflow during a certain period, in order to build an accurate model of the traffic. This model is then used for the inference of the traffic on subsequent time periods, until the model becomes inaccurate and needs to be updated. This class of methods, relying on Netflow for the calibration of a temporal model of the flows, includes but is not limited to the Kalman filtering technique, the principal component analysis, and the method of fanouts. Their common inconvenient is that the time period required for the calibration is long (at least 24 hours of measurements are needed). On the other hand, most recent methods use partial measurements of Netflow, which are collected on a regular basis, but at a limited number of locations in the network. We briefly present these methods and draw a comparative summary.

Chapter 9: Information theory and entropic projections In the information theoretic approach, we scale the vector of flow volumes so that it sums to one; the resulting vector thus

represents the distribution of probability that a packet travelling on the network belongs to a particular OD pair. Following the principle of maximum entropy, the probability distribution which best represents the current state of knowledge is, among all those distributions satisfying the measurement equations, the one with largest entropy. This gives raise to the *gravity* estimate of the traffic matrix, which is the flow distribution with maximum entropy when all we know is the volume of traffic on the external links (ingress and egress) of the network – that is, when the internal behaviour of the network is a black box.

This gravity estimate can be used as a good prior for the traffic matrix. According to Information theory, a natural approach is to select the distribution of flows which satisfies all the measurement equations (internal link counts), and is as hard to discriminate from the prior as possible (Principle of Minimum Discrimination Information). This leads to optimization problems, in which the Kullback-Leibler divergence of the flows (with respect to the gravity prior) must be minimized, subject to the constraints imposed by the linear measurements.

We next present some unpublished results on the latter optimization problem that the author obtained during his master studies. We shall see that the stationarity condition of this problem is equivalent to finding the root of a system of polynomials that is linear in every variable. We give simple conditions which ensure that a solution of this system exists and is unique. Then, we review the existing algorithms to solve this optimization problem. In particular, we analyze the similarity of the popular "Iterative proportional fitting"(IPF) algorithm with classic algorithms for matrix balancing. We shall see that the direct generalization of the matrix balancing algorithm to the case of entropic projections works if and only if all the OD pairs considered in the network are of length at most 2. In the IPF algorithm, the coordinates of the variable are updated one at a time, in a cyclic manner (instead of being updated simultaneously). This difference lets the IPF belong to the class of cyclic projection algorithms, and thus it has a linear rate of convergence.

Chapter 10: Optimization of Netflow measurements This chapter presents in greater details the results of [SBG10, SGB10]. We show that both the problem of selecting the optimal locations of Netflow and the problem of selecting the optimal sampling rates can be formulated as (linear) optimal experimental design problems. The main issue is the size of the matrices involved in this problem, which are of size $n^2 \times n^2$ for a network with n nodes. In particular, SDP approaches become intractable as soon as $n \geq 17$.

We propose a new procedure, called Successive $\boldsymbol{c}-$optimal designs (SCOD), in which we take the average of several $\boldsymbol{c}-$optimal designs, where the vectors $\boldsymbol{c}$ are drawn from a Gaussian distribution. This method can be implemented on very large networks, by solving a sequence of SOCP. Interestingly, there are some heuristic arguments which let us think that the theoretical limit of the design returned by the SCOD procedure is closed to the classic $A-$optimal design. We show by examples that this fact is verified in practice.

We next compare our SCOD approach to the greedy algorithm for the Netflow deployment problem: several networks are not (or only partially) instrumented with routers that

support Netflow. If an Internet provider wishes to equip a number of additional routers with Netflow, an interesting problem is thus to identify the most meaningful subset of locations for the monitoring-tool. Our experiments rely on real data from the *Abilene* and *Opentransit* networks (the latter is the international backbone of France Telecom).

Then, we adapt our approach so that it can take into account the past measurements (in a dynamic context, the Internet provider may not want to apply high sampling rates at the same location during successive periods of time; if a location is well measured at time t, it seems intuitive to concentrate the experimental effort to some other locations at $t+1$.) To do this, we use the ideas of a recent article of Singhal and Michailidis, in which an optimal experimental problem is stated, with an additional term in the information matrix which accounts for the errors on the past measurements, and which is computed via a Kalman filter. In fact, we shall see by an example on the Abilene network that due to the very high variability of the traffic, it is better to ignore the impact of past measurements.

Finally, we evaluate our approach for the problem of selecting the optimal sampling rates of Netflow, with per-router constraints. Given a maximal number of packets that may be sampled at each router location, the goal is to allocate optimal sampling rates to every incoming interface of each router, while keeping the number of sampled packets under the threshold. We study an instance on the Opentransit network, which contains 13456 OD pairs, 116 routers, and 436 interfaces. To the best of our knowledge, there is no other algorithm which can handle problems of this size.

Chapter 11: Perspectives for a better spatio-temporal modelling of traffic matrices We present in this chapter some perspectives for the estimation of traffic matrices. This is a preliminary work, based on the theory of random matrices and low-rank tensor decompositions.

When observed over time, the traffic matrix is a tridimensional object (origin x destination x time), which has almost always been handled as a two-dimensional object by the authors from the networking community. To this end, the Origin-Destination matrices of each time period are stacked as a column vector. By performing this vectorization though, important information on the spatial correlations between the origins and the destinations in the traffic matrix may be lost.

We have studied the empirical distribution of the singular values of the OD matrices, with the real data at our disposal from the Abilene and Opentransit backbones. Interestingly, apart from a few large singular values, the lower part of the spectrum of the OD matrices has a very good fit with the theoretical distribution of the singular values of random matrices from the so-called *Wishart* distribution. This remark lets us think that any Origin-Destination matrix can be decomposed as the sum of a low-rank matrix (which carries the *energy* of the signal), plus a noise matrix whose distribution is related to the Wishart's law. This preliminary study does not give a method for the estimation of traffic matrices from partial measurements yet. However, it sheds light on the importance of modelling the low

rank structure of OD matrices. This is done in the final section of this chapter, where low rank decompositions of the tridimensional *traffic tensor* are studied.

While low-rank approximations of matrices is a completely understood problem (through the singular value decomposition), the low-rank approximation of tensors is an active research topic. We review a few basic results and algorithms for tensor decompositions, and we show the potential of these methods by analyzing decompositions of real traffic tensors. Finally, we present the sketch of a new method –based on tensor decompositions– for the *online* estimation of traffic matrices from incomplete measurements. We show by an example on Opentransit that our method yields an improvement, by comparison to the classic *tomogravity* method.

1.4.2 Contributions of this thesis

We next list the main contributions of this thesis:

- Theorem 4.1.2, and its extension Theorem 4.2.2. Any problem from the class of *semidefinite packing programs*, where the matrix in the objective function is of rank r, has a solution of rank at most r. Extensions and consequences of this result are discussed. This theorem shall be used several times in Chapter 5.
- Theorem 5.1.1: Extension of Elfvinfg's theorem to the multiresponse case (We presented this result at the conference [SBG09]. It was discovered independently by Dette and Holland-Letz [DHL09]).
- Theorem 5.2.1: SOCP formulation of the $\boldsymbol{c}-$optimal design problem. A geometric interpretation of this result is given.
- Extension of the latter result to the case of $A-$optimality (Theorem 5.2.2), and to the case of problems with several linear inequality constraints (Theorem 5.2.3).
- Theorem 5.2.5: SOCP formulation of the $T-$optimal design problem for a subsystem of parameter $K^T\boldsymbol{\theta}$.
- Theorem 5.3.1: Geometric programming formulation of the model robust $S_\beta-$optimal design problem. The optimality conditions of this program generalize a theorem of Dette [Det93] to the case of multiresponse experiments (Theorem 5.3.2).
- A corollary of the latter result is an SOCP formulation for $D-$optimality (cf. Equation 5.25).
- Numerical tests and comparisons to other algorithms (Chapter 6), showing the importance of our SOCP approach when the number of quantities of interest r is small (typically, for $\boldsymbol{c}-$optimality where $r = 1$).
- Theorem 7.2.1: Reduction of the combinatorial *maxrank design problem* to MAX-COVERAGE. As a consequence, if $P \neq NP$, there is no polynomial-time algorithm which approximates the maxrank design by a factor larger than $1 - e^{-1}$.
- Proposition 7.2.4: If f' is operator antitone on $\mathbb{R}_+^*$, then for all triple $(X, Y, Z) \in \mathbb{S}_m^+$

$$\text{trace } f(X + Y + Z) + \text{trace } f(Z) \leq \text{trace } f(X + Z) + \text{trace } f(Y + Z).$$

- Corollary 7.2.6: The Kiefer's Φ_p-criterion (seen as a set function) is *nondecreasing submodular* for $p \in [0, 1]$.
- Theorem 7.2.7: A a consequence, the greedy algorithm always return a solution within $1-e^{-1}$ of the optimum of the Φ_p-optimal design problem ($p \in [0, 1]$). Some possible extensions of this theorem are presented.
- Proposition 7.3.4 (cf. also Theorem 2.4.7): Multiresponse generalization of the upper bound for $D-$optimal weights (discovered independently by Harman and Trnovská [HT09], for the case in which the full vector of parameters $\boldsymbol{\theta}$ is of interest, i.e. $K = \boldsymbol{I}$).
- Theorem 7.3.7: If n experiments are to be selected out of s, we present two randomized rounding algorithms which return a solution within n/s of the maxrank optimum (in average).
- We show some cases in which the performance of the latter randomized algorithms is $n/(s-1)$ (cf. Remark 7.3.2).
- Proposition 9.5.7: For the entropic projection problem with linear constraints on a network, the natural fix-point algorithm is nonexpansive if and only if every OD pair of the network is of length at most 2. (This result was obtained during the master studies of the author.)
- Optimal experimental design formulation of the Netflow deployment problem, and the Netflow optimal sampling problem (cf. Section 10.2).
- Proposition of a new method (called *Successive $\boldsymbol{c}-$Optimal Designs*, SCOD) funded on the computation of several $\boldsymbol{c}-$optimal designs to handle a certain class of large scale optimal experimental design problems (cf. Section 10.4.1). Sketch of a heuristic justification of our approach (Sections 10.4.2 and 10.4.3).
- Validation of our approach with experimental tests relying on real data (cf. Section 10.5).
- Evidence of the low-rank structure of origin-destination traffic matrices (at a given point in time). Proposition of a signal + noise model, preliminary analysis of the noise relying on the theory of random matrices (cf. Section 11.1).
- Evidence of the low rank structure of the three-way *traffic tensor* (origin $\times$ destination $\times$ time). Sketch of a method relying on tensor to estimate traffic matrices in real time (cf. Section 11.2.3).

Part I

Optimal Design of Experiments

Chapter 2

An introduction to the theory of Optimal Experiments

In this chapter, we introduce the theory of optimal experimental design, and we review the fundamental results which will be useful for the rest of this thesis.

2.1 History

The theory of optimal experimental designs has been developed since the 1920's, after some work of Gosset [Stu17] (known under the pseudonym "Student") and Fisher, who introduced several useful concepts for a theoretical approach to the design of experiments in his book [Fis35]. We refer the reader to the article of Atkinson and Bailey [AB01] for a review on the early development of the theory of optimal experiments.

One of the earliest theoretical results was obtained by Elfving in 1952 [Elf52], who focused on the problem where the experimenter disposes of s experiments, the outcome of which are linear functions of an unknown parameter (up to a zero-mean noise on the measurements). Elfving interested himself in the problem of optimally allocating a total number of n observations to the potential experiments, i.e. to select the numbers n_i of times that a measurement will be performed with experiment i, with $\sum_{i=1}^{s} n_i = n$. An idea of Elfving has been to replace the discrete design variables n_i by the real numbers $w_i = \frac{n_i}{n}$ which satisfy:

$$w_i \geq 0, \qquad \sum_{i=1}^{s} w_i = 1, \tag{2.1}$$

and then to drop the integer constraint on nw_i. In other word, Elfving posed the problem of finding the optimal amount of experimental effort w_i to spend on each experiment, where $\boldsymbol{w}$ is any continuous vector on length s satisfying Condition (2.1). A lot of results have emerged from this smoothness, starting with Elfving's Theorem (Theorem 2.4.1, [Elf52]) which characterizes geometrically the optimal design $\boldsymbol{w}$, when there is a single quantity of interest ($\boldsymbol{c}-$optimal design).

This setting was then generalized, to allow the experiments to be selected in a compact region $\mathcal{X}$, and the the design variable has become a probability measure ξ on $\mathcal{X}$. The power of this generalization was revealed in the proof of the Kiefer-Wolfowitz Theorem (Theorem 2.4.2, [KW60]), which establishes the equivalence between the two popular $D-$ and $G-$ optimality criteria.

This theorem gave birth to a sequential algorithm for the computation of $D-$optimal designs, simultaneously discovered by Wynn [Wyn70] and Fedorov [Fed72] (see Section 3.1), who further generalized the theory of optimal designs to the case of *multiresponse experiments*, where a single experiment is allowed to produce several uncorrelated observations.

Many of the optimality criteria that have been introduced for the design of experiments (including the aforementioned $\boldsymbol{c}-$ and $D-$ criteria, as well as the popular $E-$,$A-$, and $T-$ criteria which we will describe in Section 2.3.2) are convex functions of the design variable $\boldsymbol{w}$ (or ξ), and are encompassed in the class of Φ_p-criteria introduced by Kiefer [Kie75]. The work of Silvey and Titterington [ST73] and Kiefer [Kie74] showed that the Kiefer-Wolfowitz theorem could be seen as a consequence of the strong Lagrangian duality theory for convex optimization problems. Later, this result was generalized by Pukelsheim [Puk80], who established a duality theorem for a very wide class of criteria which includes the Kiefer's Φ_p- criteria.

For more details on the development of the theory of optimal experimental designs, the reader is referred to the book of Pukelsheim [Puk93].

2.2 Notation and preliminaries

2.2.1 Some notation

Throughout this thesis, we denote vectors by boldface letters and matrices by capital letters. We use the standard notation $[n] := \{1, \ldots, n\}$. The elements of a vector $\boldsymbol{x} \in \mathbb{R}^n$ are $x_1, x_2, \ldots, x_n$. The (i,j)-element of a matrix M is denoted $M_{i,j}$. The L_p-norm of the vector $\boldsymbol{x} \in \mathbb{R}^n$ is $\|\boldsymbol{x}\|_p := \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$. We shall simply denote the Euclidean norm $\|\cdot\|_2$ by $\|\cdot\|$. The vector of all zeros is written $\mathbf{0}$; similarly $\mathbf{1}$ stands for the vector of all ones. Vector inequalities should be understood elementwise, e.g. $\boldsymbol{x} \geq \mathbf{0}$ indicates that every component of $\boldsymbol{x}$ is nonnegative. The symbol ^{T} denotes the transposition operation.

The identity matrix of size $n \times n$ is denoted by $\boldsymbol{I}_n$, or simply $\boldsymbol{I}$ when there is no ambiguity. We denote by $\mathrm{Diag}(\boldsymbol{x})$ the diagonal matrix with the elements of the vector $\boldsymbol{x}$ on its diagonal, and by $\mathbf{diag}(M)$ the vector containing the diagonal entries of M. The range and nullspace of a matrix M are respectively denoted by $\mathrm{Im}\ M := \{\boldsymbol{x} :\ \exists \boldsymbol{y} : M\boldsymbol{y} = \boldsymbol{x}\}$ and $\mathrm{Ker}\ M := \{\boldsymbol{x} :\ M\boldsymbol{x} = \mathbf{0}\}$. We denote by $\mathbb{S}_m$ the space of symmetric $m \times m$ matrices.

This space is equipped with the inner product

$$\langle A, B\rangle = \mathrm{trace}(A^T B) = \sum_{i,j} a_{ij} b_{ij},$$

which induces the Frobenius norm $\|A\|_F = \sqrt{\langle A, A\rangle} = \sqrt{\sum_{i,j} a_{ij}^2}$. We also denote by $\mathbb{S}_m^+ \subset \mathbb{S}_m$ the cone of $m \times m$ symmetric positive semidefinite matrices, and by $\mathbb{S}_m^{++}$ its interior, which consists of positive definite matrices. The space of symmetric matrices is equipped with the *Löwner ordering*, which is defined by

$$\forall B, C \in \mathbb{S}_m, \qquad B \succeq C \Longleftrightarrow B - C \in \mathbb{S}_m^+. \tag{2.2}$$

Similarly, the notation $B \succ C$ indicates that $B - C$ is positive definite.

We denote by $M^\dagger$ the Moore-Penrose pseudo-inverse of M, and by M^- a *generalized inverse* of M, i.e. any matrix G verifying $MGM = M$. The reader can verify that the matrix $K_1^T M^- K_2$ does not depend on the choice of the generalized inverse when the columns of K_1 and K_2 are included in the range of M.

The convex hull (resp. conic hull) of a set $\mathcal{S}$ is denoted by $\mathrm{conv}(\mathcal{S})$ (resp. $\mathrm{cone}(\mathcal{S})$). The orthogonal of a set $\mathcal{S}$ is $\boldsymbol{S}^\perp := \{\boldsymbol{x} : \ \forall \boldsymbol{v} \in \mathcal{S}, \boldsymbol{x}^T \boldsymbol{v} = 0\}$.

2.2.2 The linear model

The most common model in optimal experimental design assumes that each experiment provides a measurement which is a linear combination of the parameters up to the accuracy of the measurement. In this thesis, we deal with linear models only. [1]

Let $\mathcal{X}$ denote the set of available experiments. Every experiment $\boldsymbol{x} \in \mathcal{X}$ provides a (multidimensional) observation

$$\boldsymbol{y}(\boldsymbol{x}) = A(\boldsymbol{x})\boldsymbol{\theta} + \boldsymbol{\epsilon}(\boldsymbol{x}), \qquad \mathbb{E}[\boldsymbol{\epsilon}(\boldsymbol{x})] = \mathbf{0} \tag{2.3}$$

where $\boldsymbol{\theta}$ is the $m-$dimensional vector of unknown parameters,

$A(\boldsymbol{x})$ is a $(l(\boldsymbol{x}) \times m)$ observation matrix, and $\boldsymbol{\epsilon}(\boldsymbol{x})$ is a zero-mean noise on the measurements with a known diagonal covariance matrix $\Sigma(\boldsymbol{x})$. The number of simultaneous observations that are collected when a measurement is performed at $\boldsymbol{x}$ is $l(\boldsymbol{x}) \leq l$. To alleviate the notation, we shall eventually write that all the observation matrices $A(\boldsymbol{x})$ are of size $l \times m$. We may always reduce to this case by setting $l - l(\boldsymbol{x})$ rows of $A(\boldsymbol{x})$ to $\mathbf{0}^T$. The mapping $\mathcal{X} \ni \boldsymbol{x} \mapsto A(\boldsymbol{x}) \in \mathbb{R}^{l \times m}$ is supposed to be continuous over $\mathcal{X}$. Note that this setting includes the common case where $\mathcal{X}$ is finite, of cardinality s, equipped with the

1. We point out that there is a theory of optimal experiments for nonlinear models, in which the design criteria depends on the unknown parameters. The basic idea is thus to search for a *locally optimal design*, which minimizes a criterion from the linear theory, for a linearization of the model at a point which is the best guess of the unknown parameters.

discrete topology, in which case we associate $\mathcal{X}$ with $[s]$ and the observation matrices are simply denoted by $A_1, \dots, A_s$.

We will assume without loss of generality that the noises have unit variance: $\Sigma(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{\epsilon}(\boldsymbol{x})\boldsymbol{\epsilon}(\boldsymbol{x})^T] = \boldsymbol{I}$. We may always reduce to this case after a left diagonal scaling of the observation equations (2.3). The errors on the measurements are assumed to be mutually independent, i.e.

$$\forall \boldsymbol{x_1} \neq \boldsymbol{x_2} \in \mathcal{X} \Longrightarrow \mathbb{E}[\boldsymbol{\epsilon}(\boldsymbol{x_1})\boldsymbol{\epsilon}(\boldsymbol{x_2})^T] = 0.$$

Uncorrelated experiments are chosen at $\boldsymbol{x_1}, \dots, \boldsymbol{x_s}$ from the experimental region $\mathcal{X}$, and the objective is to determine both the optimal choice of the $\boldsymbol{x_i}$, and the number of experiments n_i to be conducted at $\boldsymbol{x_i}$; we call such a subset of experiments a *design*. As mentioned at the beginning of this chapter, it has been proposed to work with *approximate designs*, which is simply done by releasing the integer constraints on the n_i. In this setting, a mass indicates the proportion from the total number of experiments to be conducted for each available experiment. For example, if the weight for the i^{th} experiment is w_i, and that n experiments are allowed, nw_i are chosen at $\boldsymbol{x_i}$, which suggests that each quantity nw_i is integer. However, this continuous relaxation proved to be very useful and we shall only consider approximate designs until Chapter 7, where we will focus on some combinatorial problems arising in optimal experimental design.

The design where the *percentage of experimental effort* at $\boldsymbol{x_k}$ is w_k is written as

$$\xi = \begin{pmatrix} \boldsymbol{x_1} & \cdots & \boldsymbol{x_s} \\ w_1 & \cdots & w_s \end{pmatrix},$$

or $\xi = \{\boldsymbol{x_k}, w_k\}$ for short. The set of points $\{\boldsymbol{x_i} \in \mathcal{X} : w_i > 0\}$ is called the support of ξ and is denoted by $\operatorname{supp}(\xi)$.

When $n_i = nw_i$ experiments are conducted at $\boldsymbol{x_i}$, we denote by $\overline{\boldsymbol{y}}(\boldsymbol{x_i})$ the average of these observations: we have $\mathbb{E}[\overline{\boldsymbol{y}}(\boldsymbol{x_i})] = A(\boldsymbol{x_i})\boldsymbol{\theta}$, and $\operatorname{Var}(\overline{\boldsymbol{y}}(\boldsymbol{x_i})) = \frac{1}{n_i}\boldsymbol{I}$. For the design $\xi = \{\boldsymbol{x_k}, w_k\}$, we denote by $\boldsymbol{y}(\xi)$ the aggregate vector of observations:

$$\mathbb{E}[\boldsymbol{y}(\xi)] = A(\xi)\ \boldsymbol{\theta}, \tag{2.4}$$

$$\text{where} \qquad \boldsymbol{y}(\xi) = \begin{pmatrix} \overline{\boldsymbol{y}}(\boldsymbol{x_1}) \\ \vdots \\ \overline{\boldsymbol{y}}(\boldsymbol{x_s}) \end{pmatrix}, \qquad \text{and} \quad A(\xi) = \begin{bmatrix} A(\boldsymbol{x_1}) \\ \vdots \\ A(\boldsymbol{x_s}) \end{bmatrix}.$$

In addition, the variance of this aggregate observation vector satisfies $\operatorname{Var}(\boldsymbol{y}(\xi)) = \frac{1}{n}\Delta(\boldsymbol{w})$, where

$$\Delta(\boldsymbol{w}) = \begin{pmatrix} 1/w_1\boldsymbol{I} & & \\ & \ddots & \\ & & 1/w_s\boldsymbol{I} \end{pmatrix}, \tag{2.5}$$

with $(l(\boldsymbol{x_i}) \times l(\boldsymbol{x_i}))-$identity blocks on the diagonal. If $w_i = 0$ for some $i \in [s]$, we simply remove the measurement point $\boldsymbol{x_i}$ from ξ. For ease of presentation, we get rid of the

multiplication factor $1/n$, since it does not affect the results on optimal designs.

2.2.3 Gauss-Markov Theorem and Information matrices

The linear theory assumes that the experimenter is interested in estimating the vector

$$\boldsymbol{\zeta} = K^T\boldsymbol{\theta},$$

where K is of size $m \times r$ and has full column rank. In other words, the experimenter wants to estimate a collection $(\zeta_1, \dots, \zeta_r)$ of linear combinations of the parameters. We denote the columns of K by $\boldsymbol{c_1}, \dots, \boldsymbol{c_r}$, so that the quantities of interest are:

$$\forall i \in [r], \quad \zeta_i = \boldsymbol{c_i}^T\boldsymbol{\theta}.$$

This setting includes the cases $K = \boldsymbol{I}$, in which the experimenter wants to estimate each individual parameter θ_i, and the case $r = 1$ (known as $\boldsymbol{c}-$optimality in the literature) in which there is a single quantity of interest $\zeta = \boldsymbol{c}^T\boldsymbol{\theta}$.

It can easily be seen that a linear estimator $\hat{\boldsymbol{\zeta}} = H^T\boldsymbol{y}(\xi)$ is unbiased if and only if $A(\xi)^T H = K$. Thus, linear unbiased estimators for $\boldsymbol{\zeta}$ exist as long as the columns of K are in the range of $A(\xi)^T$. In the sequel, we will say that the vector $\boldsymbol{\zeta} = K^T\boldsymbol{\theta}$ is *estimable* if there exists a design ξ such that the latter condition is satisfied. Notice that a sufficient condition which ensures that $K^T\boldsymbol{\theta}$ is estimable for any $m \times r$ matrix K is that the matrices $\Big(A(\boldsymbol{x})\Big)_{\boldsymbol{x}\in\mathcal{X}}$ contain m linearly independent vectors among their rows. For an estimable quantity $K^T\boldsymbol{\theta}$, we define the feasibility cone $\Xi(K)$ as the set of designs ξ such that $A(\xi)^T$ span the columns of K, and a design ξ will be said *feasible* if it lies in the feasibility cone.

We are interested in finding the *best* unbiased estimator for $\boldsymbol{\zeta}$, in the sense that its variance should be minimal. The variance of a vector is in fact a positive semidefinite matrix, and so the comparison between two covariance matrices should be in terms of Löwner ordering (cf. Page 29). The *Gauss-Markov* theorem, which is a classical result in the field of statistics, gives the form of this best estimator. We give below a proof of this theorem relying on the Schur complement lemma.

Theorem 2.2.1 (Gauss-Markov Theorem)**.** *Let $K^T\boldsymbol{\theta}$ be estimable and $\xi = \{\boldsymbol{x_k}, w_k\} \in \Xi(K)$ be a feasible design. For any matrix H such that $A(\xi)^T H = K$, $\hat{\boldsymbol{\zeta}} = H^T\boldsymbol{y}(\xi)$ is an unbiased estimator for $\boldsymbol{\zeta}$, and its covariance matrix satisfies*

$$Var(\hat{\boldsymbol{\zeta}}) = H^T \operatorname{Var}(\boldsymbol{y}(\xi))H = H^T\Delta(\boldsymbol{w})H \succeq K^T\left(A(\xi)^T\Delta(\boldsymbol{w})^{-1}A(\xi)\right)^- K.$$

Moreover, this latter bound is attained for the estimator $\hat{\boldsymbol{\zeta}}^* = H^{*T}\boldsymbol{y}(\xi)$, where

$$H^* = \Delta(\boldsymbol{w})^{-1}A(\xi)(A(\xi)^T\Delta(\boldsymbol{w})^{-1}A(\xi))^\dagger K. \tag{2.6}$$

Proof. The fact that the lower bound is attained for $\hat{\boldsymbol{\zeta}}^* = H^{*T}\boldsymbol{y}(\xi)$ is clear by substituting H^* to H in the expression of the variance of $\hat{\boldsymbol{\zeta}}$, and by using the fact that for any matrix M, we have $M^\dagger M M^\dagger = M^\dagger$.

Hence, the only thing to prove is the matrix inequality. The matrix

$$\begin{pmatrix} A(\xi)^T\Delta(\boldsymbol{w})^{-1}A(\xi) & K \\ K^T & H^T\Delta(\boldsymbol{w})H \end{pmatrix}$$

is positive semidefinite, because it can be written as the following product:

$$\begin{pmatrix} A(\xi)^T\Delta(\boldsymbol{w})^{-1/2} \\ H^T\Delta(\boldsymbol{w})^{1/2} \end{pmatrix} \begin{pmatrix} \Delta(\boldsymbol{w})^{-1/2}A(\xi) & \Delta(\boldsymbol{w})^{1/2}H \end{pmatrix}.$$

The Schur complement lemma indicates that since $H^T\Delta(\boldsymbol{w})H \succeq 0$, the matrix

$$H^T\Delta(\boldsymbol{w})H - K^T\left(A(\xi)^T\Delta(\boldsymbol{w})^{-1}A(\xi)\right)^- K$$

must be positive semidefinite. This completes the proof. □

Remark 2.2.1. An alternative formulation of the Gauss-Markov Theorem states that if Σ is nonsingular and the columns of K are in the range of the matrix A^T, then the optimization problem

$$\begin{aligned} \min_H{}_{\preceq} \quad & H^T\Sigma H \\ \text{s.t.} \quad & A^T H = K, \end{aligned}$$

where the minimum is taken with respect to the Löwner ordering, attains its solution for $H = \Sigma^{-1}A(A^T\Sigma^{-1}A)^\dagger K$, and the value of the *minimum* is $K^T(A^T\Sigma^{-1}A)^- K$.

Gauss Markov theorem gives the form of the *best unbiased linear estimator*, and shows that its variance is

$$\mathrm{Var}(\hat{\boldsymbol{\zeta}}^*) = K^T(A(\xi)^T\Delta(\boldsymbol{w})^{-1}A(\xi))^- K = K^T M(\xi)^- K, \tag{2.7}$$

where $M(\xi)^-$ is a *generalized inverse* of $M(\xi) := A(\xi)^T\Delta(\boldsymbol{w})^{-1}A(\xi)$ and the reader can verify that the latter expression does not depend on the choice of the generalized inverse. The positive semidefinite matrix $M(\xi)$ is called the *information matrix* of the design. We also define the partial information matrices of each experiment $M(\boldsymbol{x}) := A(\boldsymbol{x})^T A(\boldsymbol{x})$, so that $M(\xi)$ can be decomposed as a weighted sum of the information matrices of the selected experiments:

$$M(\xi) = \sum_{i=1}^{s} w_i A(\boldsymbol{x_i})^T A(\boldsymbol{x_i}) = \sum_{i=1}^{s} w_i M(\boldsymbol{x_i}) \tag{2.8}$$

Remark 2.2.2. If we further assume that the noise follows a normal distribution $\mathcal{N}(0, \boldsymbol{I})$, then the estimator $\hat{\boldsymbol{\zeta}}^*$ described in (2.6) is also the maximum likelihood estimator of $\boldsymbol{\zeta}$, and

the bound given by the Cramer-Rao inequality is attained, i.e. its covariance matrix equals the inverse of the Fisher information matrix.

We next define the $K-$information matrix $Q_K(\xi) = (K^T M(\xi)^- K)^{-1}$ as the inverse of the covariance matrix[2]. Note that the inverse is well defined when $\xi \in \Xi(K)$. Otherwise, it is still possible to extend the definition of $Q_K(\xi)$ per continuity; in fact, the correct definition of the $K-$information matrix is given in Chapter 3 of [Puk93]:

$$\begin{aligned} Q_K(\xi) := \quad & \min_{L} \preceq \quad L^T M(\xi) L \\ & \text{s.t.} \quad K^T L = \boldsymbol{I}_r, \end{aligned} \tag{2.9}$$

where the minimum is taken with respect to Löwner ordering. Pukelsheim shows that the minimum exists indeed (which is not obvious since the Löwner ordering is a partial ordering), as a consequence of the Gauss-Markov Theorem (cf. Theorem 1.21 in [Puk93]). In the sequel, the reader needs only remind the simple expression $Q_K(\xi) = (K^T M(\xi)^- K)^{-1}$, which is valid in the regular case $\xi \in \Xi(K)$, and that the matrix $Q_K(\xi)$ exists and is singular when the range of $M(\xi)$ does not include the range of K (that is, when $\xi \notin \Xi(K)$).

The reader may wonder why we reduce ourselves to the case of designs with a finite –or even countable– number of support points. It was proposed indeed to work in a more general framework, by allowing the design to take the form of a probability measure ω over the regression region $\mathcal{X}$, so that the information matrix becomes

$$M(\omega) = \int_{\mathcal{X}} A(\boldsymbol{x})^T A(\boldsymbol{x}) d\omega(\boldsymbol{x}).$$

However, this continuous form of the information matrix is still a symmetric matrix from the closed convex hull of $\{A(\boldsymbol{x})^T A(\boldsymbol{x}),\ \boldsymbol{x} \in \mathcal{X}\}$. When $\mathcal{X}$ is compact, and $\boldsymbol{x} \mapsto A(\boldsymbol{x})$ is continuous, the set of all information matrices $\{A(\boldsymbol{x})^T A(\boldsymbol{x}),\ \boldsymbol{x} \in \mathcal{X}\}$ is closed, and we know from Caratheodory's theorem that $M(\omega)$ can be written as barycenter of $m(m+1)/2+1$ information matrices (see Fedorov [Fed72]). Therefore, the optimal design can always be expressed with a discrete measures $\omega = w_1 \delta(\boldsymbol{x} - \boldsymbol{x_1}) + ... + w_s \delta(\boldsymbol{x} - \boldsymbol{x_s})$, where $s \leq m(m+1)/2+1$, and we will consider only such designs in this work. Moreover, the study of designs with a discrete support is appropriate for the framework of the industrial application of the second part of this thesis.

2.3 Optimality criteria

2.3.1 c-optimality

The experimental design approach consists in choosing the design ξ in order to make the variance of the estimator (2.7) *as small as possible*. The problem is well posed when

2. Note that If $K = \boldsymbol{I}$ (i.e. when the experimenter wants to estimate the whole parameter $\boldsymbol{\theta}$), then the $K-$information matrix $Q_K(\xi)$ coincides with the information matrix $M(\xi)$.

$r = 1$, since in this case the variance is a scalar. This is the framework for the $\boldsymbol{c}-$optimal design problem, in which K has a single column $\boldsymbol{c}$, and the problem is now to find the design $\xi = \{\boldsymbol{x_k}, w_k\}$ minimizing the variance (2.7):

$$\begin{aligned} \min_{\xi=\{\boldsymbol{x_k},w_k\}\in\Xi(\boldsymbol{c})} \quad & \boldsymbol{c}^T M(\xi)^- \boldsymbol{c} \\ \text{s.t.} \quad & M(\xi) = \sum_{i=1}^{s} w_i A(\boldsymbol{x_i})^T A(\boldsymbol{x_i}) \\ & \sum_{i=1}^{s} w_i = 1, \quad \forall\, i \in [s],\ w_i \geq 0, \boldsymbol{x_i} \in \mathcal{X}. \end{aligned} \tag{2.10}$$

This problem was first studied by Elfving, in the case of *single response experiments*, i.e. when each experiment yields only one observation ($\forall \boldsymbol{x} \in \mathcal{X},\ l(\boldsymbol{x}) = 1$ and $A(x)$ is a row vector.) In his pioneer work, Elfving discovered a geometrical characterization of $\boldsymbol{c}-$optimality [Elf52] which we will detail in Section 2.4.1.

2.3.2 The class of Kiefer's Φ_p criteria

When $r > 1$, the natural problem is to minimize the covariance matrix of the best linear unbiased estimator (2.7) with respect to the Löwner ordering. A geometrical interpretation of this problem is the following: with the assumption that the noise $\boldsymbol{\epsilon}(\boldsymbol{x})$ is normally distributed for all $\boldsymbol{x} \in \mathcal{X}$, for every probability level α, the best estimator $\hat{\boldsymbol{\zeta}}^*$ lies in the confidence ellipsoid centered at $\boldsymbol{\zeta}$ and defined by the following inequality:

$$(\boldsymbol{\zeta} - \hat{\boldsymbol{\zeta}}^*)^T Q_K(\xi)(\boldsymbol{\zeta} - \hat{\boldsymbol{\zeta}}^*) \leq \kappa_\alpha, \tag{2.11}$$

where κ_α depends on the specified probability level. We would like to make these confidence ellipsoids *as small as possible*, in order to reduce the uncertainty on the estimation of $\boldsymbol{\zeta}$. To this end, we can express the inclusion of ellipsoids in terms of matrix inequalities. One can readily check that for any value of the probability level α, the confidence ellipsoid (2.11) corresponding to $Q_K(\xi)$ is included in the confidence ellipsoid corresponding to $Q_K(\xi')$ if and only if $Q_K(\xi) \succeq Q_K(\xi')$. Hence, we will prefer design ξ to design ξ' if the latter inequality is satisfied, and we want to select a design which maximizes $Q_K(\xi)$ (or equivalently which minimizes its inverse $K^T M(\xi)^- K$) for the Löwner ordering.

Since Löwner ordering is only a partial ordering on $\mathbb{S}_m$ (and the inclusion relation is a partial ordering on the ellipsoids of $\mathbb{R}^m$), the problem consisting in maximizing $Q_K(\xi)$ is ill-posed. Hence, we will rather maximize a scalar *information function* of the $K-$information matrix, i.e. a function mapping $\mathbb{S}_m^+$ onto the real line, and which satisfies natural properties, as positive homogeneity, monotonicity with respect to Löwner ordering, and concavity. Kiefer [Kie75] proposed to make use of the class of matrix means Φ_p. These functions are defined like the L_p-norm of the vector of eigenvalues of the information matrix, but for $p \in [-\infty, 1]$. For positive definite matrices, $M \in \mathbb{S}_m^{++}$ with eigenvalues $\{\lambda_1, \ldots, \lambda_m\}$, the matrix mean Φ_p is defined by

$$\Phi_p(M) = \begin{cases} \lambda_{\min}(M) & \text{for } p = -\infty \ ; \\ (\frac{1}{m} \text{ trace } M^p)^{\frac{1}{p}} & \text{for } p \in]-\infty, 1], \ p \neq 0; \\ (\det(M))^{\frac{1}{m}} & \text{for } p = 0, \end{cases} \tag{2.12}$$

where we have used the extended definition of powers of matrices M^p for arbitrary real parameters p: $\text{trace } M^p = \sum_{j=1}^m \lambda_j^p$. For singular positive semidefinite matrices, Φ_p is defined by continuity:

$$\Phi_p(M) = \begin{cases} 0 & \text{for } p \in [-\infty, 0] \ ; \\ (\frac{1}{m} \text{ trace } M^p)^{\frac{1}{p}} & \text{for } p \in]0, 1]. \end{cases} \tag{2.13}$$

The reader is referred to Pukelsheim [Puk93] for a complete analysis of these information functions. For a real $p \in [-\infty, 1]$, the problem of Φ_p-optimality is

$$\begin{aligned} \max_{\xi=\{\boldsymbol{x_k}, w_k\} \in \Xi(K)} \quad & \Phi_p\Big(Q_K(\xi)\Big) \\ \text{s.t.} \quad & \sum_{i=1}^s w_i = 1, \quad \forall \ i \in [s], \ w_i \geq 0, \boldsymbol{x_i} \in \mathcal{X}. \end{aligned} \tag{2.14}$$

This class of problems was introduced by Kiefer in 1975, and it interpolates several popular criteria which were used long before. We next review these criteria, which are obtained for special value of p. A remarkable property of these optimization problems is that, if $K^T\boldsymbol{\theta}$ is estimable, and except for the pathological case $p = 1$, the constraint $\xi \in \Xi(K)$ can be removed without changing the optimum. The extended feasible space

$$\{\xi = \{\boldsymbol{x_k}, w_k\}, \quad \forall \ i \in [s], \ \boldsymbol{x_i} \in \mathcal{X} w_i \geq 0; \quad \sum_{i=1}^s w_i = 1\} \tag{2.15}$$

is compact, which guarantees the existence of an optimal solution ξ^* (because the objective function is continuous). This fundamental existence result is presented in a unified way for Kiefer's Φ_p-criteria ($p < 1$) in [Puk93]. Following Pukelsheim's terminology, we call a design *formally* $\Phi-$*optimal* if it maximizes $\Phi(Q_K(\xi))$ in the set (2.15). The estimability of $K^T\boldsymbol{\theta}$ implies that there is a design ξ such that $Q_K(\xi)$ is nonsingular. Now, for all $p \leq 0$, the Φ_p-criterion vanishes for singular matrices. It follows that any formally Φ_p-optimal design ξ is such that $\Phi_p(Q_K(\xi)) > 0$. Recall that the definition of $Q_K(\xi)$ can be extended to the designs that are not feasible, and for which $K^T M(\xi) K$ fails to be invertible (see the discussion following Equation (2.9)). The key point is that $Q_K(\xi)$ becomes singular when $\xi \notin \Xi(K)$. Hence, the optimal design $\xi \in \Xi(K)$ and solves Problem (2.14). For all $p \in]0, 1[$, a similar argument holds, by considering the Fenchel conjugate function $m\Phi_q$ of Φ_p (here, q is the real number such that $\frac{1}{p} + \frac{1}{q} = 0$, see Section 7.13 in Pukelsheim [Puk93]).

D-Optimality

The $D-$criterion is obtained for $p = 0$, and consists in maximizing the determinant of the $K-$information matrix:

$$\begin{aligned} \max_{\xi=\{\boldsymbol{x_k},w_k\}} \quad & \det\Big(Q_K(\xi)\Big) \\ \text{s.t.} \quad & \sum_{i=1}^{s} w_i = 1, \quad \forall\ i \in [s],\ w_i \geq 0, \boldsymbol{x_i} \in \mathcal{X}. \end{aligned} \tag{2.16}$$

We have seen above that the maximization of $Q_K(\xi)$ with respect to the Löwner ordering was equivalent to the minimization of any ellipsoid of the form (2.11) for the inclusion relation. In fact, such ellipsoids have their axis aligned with the eigenvectors of $Q_K(\xi)$, and the semi-axis in the direction of the eigenvector associated with the eigenvalue λ_i is of length proportional to $\frac{1}{\sqrt{\lambda_i}}$. This allows a nice geometrical interpretation of this criterion: The volume of the ellipsoid (2.11) is given by $C_m \kappa_\alpha^{m/2} \det(Q_K(\xi))^{-1/2}$ where $C_m > 0$ is a constant depending only on the dimension. Therefore, the $D-$optimal design minimizes the volume of the ellipsoids (2.11), which coincide with the confidence ellipsoids of $\hat{\boldsymbol{\zeta}}^*$ in the Gaussian case (cf. Figure 2.1**(a)**).

E-Optimality

The $E-$criterion is obtained for $p = -\infty$. It consists in maximizing the smallest eigenvalue of $\Big(Q_K(\xi)\Big)$.

$$\begin{aligned} \max_{\xi=\{\boldsymbol{x_k},w_k\}} \quad & \lambda_{\min}\Big(Q_K(\xi)\Big) \\ \text{s.t.} \quad & \sum_{i=1}^{s} w_i = 1, \quad \forall\ i \in [s],\ w_i \geq 0, \boldsymbol{x_i} \in \mathcal{X}. \end{aligned} \tag{2.17}$$

As for the $D-$criterion, we can give a geometrical interpretation to this criterion: the $E-$optimal design minimizes the length of the largest semi-axis of the ellipsoids (2.11), as plotted on Figure 2.1**(b)**.

A-Optimality

The $A-$criterion is obtained for $p = -1$, and aims at maximizing the harmonic average of the eigenvalues of the $K-$information matrix, or equivalently at minimizing its inverse:

$$\begin{aligned} \min_{\xi=\{\boldsymbol{x_k},w_k\}} \quad & \frac{1}{m} \text{trace } Q_K(\xi)^{-1} \\ \text{s.t.} \quad & \sum_{i=1}^{s} w_i = 1, \quad \forall\ i \in [s],\ w_i \geq 0, \boldsymbol{x_i} \in \mathcal{X}. \end{aligned} \tag{2.18}$$

If we denote the eigenvalues of $Q_K(\xi)^{-1} = K^T M(\xi)^- K$ by $\lambda_1, \ldots, \lambda_m$, this harmonic average can also be written as

$$\Phi_A(\xi) = m \sum_{i=1}^{m} \frac{1}{\lambda_i} = m \sum_{i=1}^{m} \left(\frac{1}{\sqrt{\lambda_i}} \right)^2 .$$

From this expression, we see that the $A-$optimal design minimizes the diagonal of the bounding box of the ellipsoids (2.11), as shown on Figure 2.1**(c)**.

T-Optimality

The $T-$criterion is obtained for $p = 1$, and aims at maximizing the trace of the $K-$information matrix.

$$\begin{aligned} \sup_{\xi=\{\boldsymbol{x_k},w_k\}\ \in \Xi(K)} \quad & \text{trace } Q_K(\xi) \\ \text{s.t.} \quad & \sum_{i=1}^{s} w_i = 1, \quad \forall\ i \in [s],\ w_i \geq 0, \boldsymbol{x_i} \in \mathcal{X}. \end{aligned} \tag{2.19}$$

This criterion is not much used in practice, because of its pathological behavior. Since $M \mapsto \Phi_1(M)$ is not *strictly* concave (it is linear), a formally Φ_1-optimal design ξ can fail to be feasible for problem (2.19), i.e. $\xi \notin \Xi(K)$. Moreover, we will see in Section 2.4.3 that every $T-$optimal design for the full parameter $\boldsymbol{\theta}$ is concentrated on the points $\boldsymbol{x}$ such that $\|A(\boldsymbol{x})\|_F$ is maximal, which is not a good recommendation in practice. We give below an example where Problem (2.19) has no solution, i.e. where the supremum over $\xi \in \Xi(K)$ is not attained. Consider a simple regression model with only two experiments ($\mathcal{X} = \{1, 2\}$), and row observation matrices $A_1 = [1, 0],\ A_2 = [0, 2]$. The information matrix for this model is

$$M(\xi) = \begin{pmatrix} w_1 & \\ & 4w_2 \end{pmatrix}.$$

When the full parameter is of interest ($K = \boldsymbol{I}$), the design $\boldsymbol{w}$ is feasible if and only if $M(\xi)$ is invertible, i.e. $\boldsymbol{w} > \boldsymbol{0}$. We have $Q_{\boldsymbol{I}}(\xi) = (M(\xi)^{-1})^{-1} = M(\xi)$, which remains true even for the nonfeasible designs where $w_1 = 0$ or $w_2 = 0$ by continuity of $\xi \mapsto Q_{\boldsymbol{I}}(\xi)$. The trace

of the information matrix is maximized over the set $\{\boldsymbol{w} : \ w_1 + w_2 = 1, \ \boldsymbol{w} \geq \mathbf{0}\}$ for the non-feasible design $\boldsymbol{w} = [0, 1]^T$. Furthermore, the optimal value of Problem (2.19) can be approached from below with arbitrary precision by the feasible designs $\boldsymbol{w_\epsilon} = [\varepsilon, 1 - \varepsilon]^T$, where $\epsilon \to 0^+$.

2.3.3 S-optimality: a model robust criterion

The S-criterion was introduced by Läuter [Läu74] in order to tackle the uncertainty of the experimenter on the *true model*, by considering a class of r plausible models with means

$$\mathbb{E}[\boldsymbol{y}(\boldsymbol{x})] = A_{(i),\boldsymbol{x}}\theta \qquad i \in [r],$$

in which the quantity to estimate is $\zeta_i = \boldsymbol{c_i}^T\boldsymbol{\theta}$.

In other words, the measurement $\boldsymbol{y}(\boldsymbol{x})$ at $\boldsymbol{x}$ is modeled as a linear function of the parameter $\boldsymbol{\theta}$, which depends on the model, and must be used to estimate a linear function ζ of the parameter in each model. In practice, the parameters of each of these models may be different. This can be handled by setting the j^{th} column of $A_{(i),\boldsymbol{x}}$ to $\mathbf{0}$ whenever the i^{th} model at $\boldsymbol{x}$ does not depend on θ_j. Note that we write the index of the model in parenthesis, in order to avoid ambiguities with the index of the experiment.

Given a nonnegative vector $\boldsymbol{\beta}$ of size r with sum 1, where β_i indicates the importance that the experimenter attaches to the model i, or the importance of the linear combination $\boldsymbol{c_i}^T\boldsymbol{\theta}$, the $S_\beta-$criterion is:

$$S_{\boldsymbol{\beta}}(\xi) = \sum_{i=1}^{r} \beta_i \log(\boldsymbol{c_i}^T M_{(i)}(\xi)^{-}\boldsymbol{c_i}),$$

where

$$M_{(i)}(\xi) = \sum_{k=1}^{s} w_k A_{(i),\boldsymbol{x_k}}^T A_{(i),\boldsymbol{x_k}}$$

is the information matrix in the i^{th} model. A design minimizing this criterion is called $S_\beta-$optimal. An interesting case occurs when the s models are identical. This is an alternative approach to the $A-$optimality for $K^T\boldsymbol{\theta}$, with weightings on each linear combination $\boldsymbol{c_i}^T\boldsymbol{\theta}$ to be estimated. Dette studied the difference between these two approaches in Section 4 of [Det93].

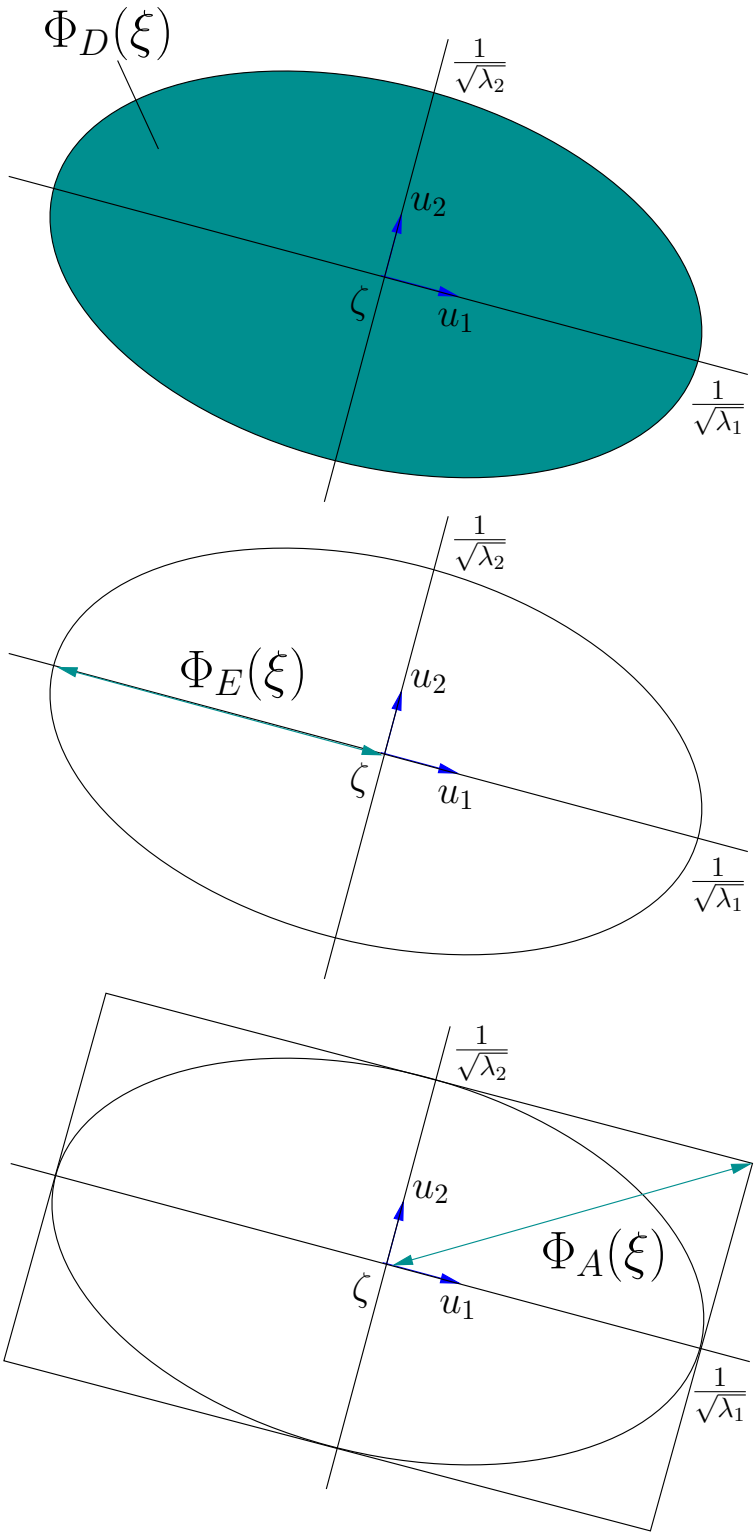


Figure 2.1: Geometrical interpretation of $D-$, $E-$ and $A-$ optimality criteria. The ellipsoids (2.11) are plotted in two dimensions, for $\kappa_\alpha = 1$ and when the $K-$information matrix has a singular value decomposition of the form $Q_K(\xi) = \lambda_1 \boldsymbol{u_1} \boldsymbol{u_1}^T + \lambda_2 \boldsymbol{u_2} \boldsymbol{u_2}^T$. The $D-$criterion **(a)** corresponds to the volume, the $E-$criterion **(b)** to the largest semi-axis and the $A-$criterion **(c)** to the diagonal of the bounding box of the ellipsoids.

2.4 Fundamental results

2.4.1 Elfving's Theorem for c-optimality

Elfving's result [Elf52] describes the geometry of $\boldsymbol{c}-$optimal designs. This is one of the earliest theoretical result in the theory of optimal design of experiments, and its importance was illustrated in many works [Che99, Det93, DS93, HHS95, Stu71, Stu05]. Elfving studied the $\boldsymbol{c}$-optimal design problem in the case of *single response experiments*[3], i.e. when each experiment yields only one observation ($\forall \boldsymbol{x} \in \mathcal{X},\ l(\boldsymbol{x}) = 1$ and $A(\boldsymbol{x})$ is a row vector which we denote by ${\boldsymbol{a}_{\boldsymbol{x}}}^T$; *beware of the transposition, we use a different convention for the observation matrix in the single response case because we prefer seeing the regression vectors as column vectors*). We will show that a generalization of Elfving's theorem to the case of multiresponse experiments is possible in Chapter 5.

We first define the *Elfving set* as the convex hull of the vectors $\pm\boldsymbol{a}_{\boldsymbol{x}}$:

$$\mathcal{E} = \operatorname{conv}\Big(\{\pm\boldsymbol{a}_{\boldsymbol{x}},\ \boldsymbol{x} \in \mathcal{X}\}\Big), \tag{2.20}$$

and we denote its boundary by $\partial\mathcal{E}$.

Theorem 2.4.1 (Elfving [Elf52])**.** *A design $\xi = \{\boldsymbol{x_i}, w_i\}$ is $\boldsymbol{c}-$optimal if and only if there exists scalars $\epsilon_i = \pm 1$ and a positive real t such that*

$$t\boldsymbol{c} = \sum_{i=1}^{s} w_i \epsilon_i \boldsymbol{a}_{\boldsymbol{x_i}} \in \partial\mathcal{E}.$$

Moreover, $t^{-2} = \boldsymbol{c}^T M(\xi)^{-}\boldsymbol{c}$ is the minimal variance.

The generalization to multiresponse experiments that we give in Section 5.1 has a proof relying on original ideas of Elfving, and so we will only prove this generalization (Theorem 5.1.1)). Elfving's theorem shows that the $\boldsymbol{c}-$optimal design is characterized by the intersection between the vectorial straight line directed by $\boldsymbol{c}$ and the boundary of the Elfving set $\mathcal{E}$. We also point out that when the vector $\boldsymbol{c}$ is not spanned by the regression vectors $(\boldsymbol{a}_{\boldsymbol{x}})_{\boldsymbol{x}\in\mathcal{X}}$, in other words when $\boldsymbol{c}^T\boldsymbol{\theta}$ is not estimable (i.e. $\Xi(\boldsymbol{c}) = \emptyset$), then the only scalar t such that $t\boldsymbol{c}$ lies in $\mathcal{E}$ is 0, and so a $\boldsymbol{c}-$optimal design does not exist, in accordance with the discussion in the second paragraph of Section 2.2.3.

We show on Figure 2.2 a representation of Elfving's theorem in dimension 2. Here, $\mathcal{X} = \{1, 2, 3, 4\}$ is finite, so that the Elfving set is a polyhedron, and we write $\boldsymbol{a_i}$ for $\boldsymbol{a}_{\boldsymbol{x_i}}$. The vector $\boldsymbol{c}$ is along the θ_1-axis, which means that the experimenter wants to estimate $\zeta = \theta_1$. The intersection between this axis and the Elfving set indicates the optimal weights of the $\boldsymbol{c}-$optimal design: $w_3 = \frac{3}{4}$ and $w_4 = \frac{1}{4}$. Note that since $\boldsymbol{a_2}$ is in the interior of the Elfving set, the experiment 2 is never selected, whatever is the vector $\boldsymbol{c}$. This example also shows that the optimal design $\boldsymbol{w}^*$ can be computed by linear programming when $\mathcal{X}$ is finite (intersection of a straight line and a polyhedron). We will study this feature in Chapter 3.

3. The more general setting of *multiresponse experiments* was introduced by Fedorov in 1972 [Fed72]

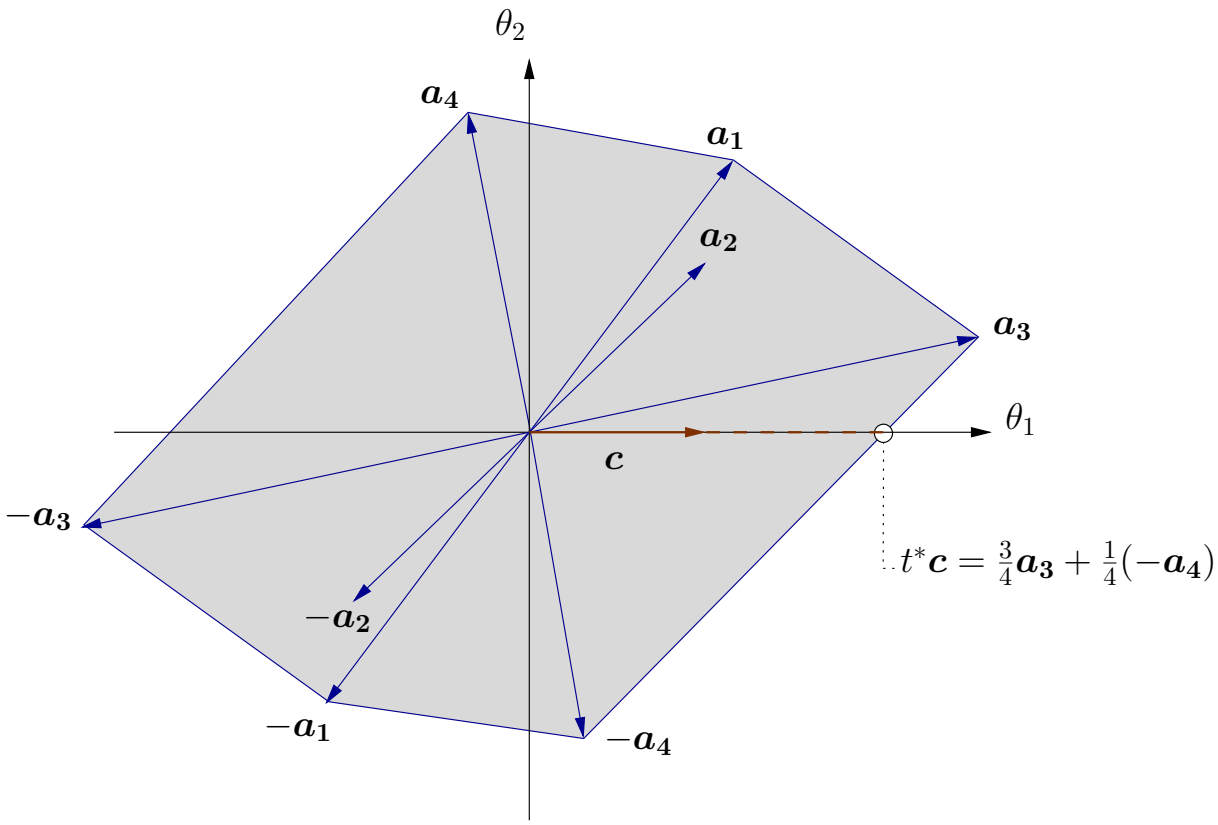


Figure 2.2: Geometrical representation of Elfving's theorem in dimension two. The grey area represents the Elfving set, which is a polyhedron because $\mathcal{X}$ is finite (here, $\mathcal{X} = \{1, 2, 3, 4\}$). The intersection $t^*\boldsymbol{c}$ determines the weights of the $\boldsymbol{c}-$optimal design: $\boldsymbol{w}^* = [0, 0, \frac{3}{4}, \frac{1}{4}]^T$.

2.4.2 The Kiefer-Wolfowitz Theorem for D-optimality

The Kiefer-Wolfowitz theorem [KW60] was established for single-response experiments in 1960, and then extended to the multiresponse framework by Fedorov [Fed72]. We give below both versions of this theorem.

A special case of $\boldsymbol{c}-$optimality is when the experimenter wants to estimate a quantity $\zeta = \boldsymbol{a_x}^T\boldsymbol{\theta}$ which can be observed by a single experiment (here, the experiment at $\boldsymbol{x}$ with regression vector $\boldsymbol{a_x}$). In this case, the variance of the best estimator is $\boldsymbol{a_x}^T M(\xi)^- \boldsymbol{a_x}$. This case is highly trivial since the experimenter's interest is to affect all the experimental effort to $\boldsymbol{x}$. However, an interesting case occurs when the experimenter is not interested in the observation of a single experiment $\boldsymbol{a_x}^T\boldsymbol{\theta}$, but in the whole *regression surface* $\{\boldsymbol{a_x}^T\boldsymbol{\theta},\ \boldsymbol{x} \in \mathcal{X}\}$. A global criterion is needed to measure the performance of a design in this case. The *global criterion* (known as $G-$criterion) is

$$\Phi_G(\xi) = \max_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{a_x}^T M(\xi)^- \boldsymbol{a_x}$$

and the $G-$optimal design guards one against the worst case, by minimizing the variance

of every observation in the regression surface:

$$\begin{aligned} \min_{\xi} \quad & \max_{\boldsymbol{x}\in\mathcal{X}} \ \boldsymbol{a_x}^T M(\xi)^- \boldsymbol{a_x} \\ \text{s.t.} \quad & M(\xi) = \sum_{i=1}^{s} w_i A(\boldsymbol{x_i})^T A(\boldsymbol{x_i}) \\ & \sum_{i=1}^{s} w_i = 1, \quad \forall\ i \in [s], w_i \geq 0, \boldsymbol{x_i} \in \mathcal{X}. \end{aligned} \tag{2.21}$$

The Kiefer-Wolfowitz theorem establishes the equivalence between the $D-$optimal design and the $G-$optimal design:

Theorem 2.4.2 (Kiefer-Wolfowitz [KW60])**.** *Assume that the regression range $(\boldsymbol{a_x})_{x\in\mathcal{X}}$ contains m linearly independent vectors. Then the following statements are equivalent:*

(i) The design ξ is $G-$optimal;

(ii) The design ξ is $D-$optimal for the full parameter $\boldsymbol{\theta}$ (i.e. with $K = \boldsymbol{I}$);

(iii) For all $\boldsymbol{x}$ in $\mathcal{X}$, $\boldsymbol{a_x}^T M(\xi)^- \boldsymbol{a_x} \leq m$.

Moreover, the bound provided by the inequality in (iii) is attained for the support points of the optimal design:

$$\boldsymbol{x_i} \in \operatorname{supp}(\xi) \Longrightarrow \boldsymbol{a_{x_i}}^T M(\xi)^- \boldsymbol{a_{x_i}} = m.$$

Proof. We first show that for all design $\xi = \{\boldsymbol{x_k}, w_k\}$, we have $\Phi_G(\xi) \geq m$. If $M(\xi)$ is singular, then by assumption there is a regression vector $\boldsymbol{a_x}$ which is not in the range of $M(\xi)$, and so $\Phi_G(\xi) = \infty \geq m$. If $M(\xi)$ is nonsingular, we write:

$$\begin{aligned} m = \operatorname{trace}\ \boldsymbol{I} = \operatorname{trace}\ M(\xi)M(\xi)^{-1} &= \operatorname{trace}(\sum_{i=1}^{s} w_i \boldsymbol{a_{x_i}} \boldsymbol{a_{x_i}}^T M(\xi)^{-1}) \\ &\leq \sum_{i=1}^{s} w_i \ \max_{\boldsymbol{x}\in\mathcal{X}} (\boldsymbol{a_x}^T M(\xi)^{-1} \boldsymbol{a_x}) \\ &= \Phi_G(\xi). \end{aligned}$$

This proves the part $(iii) \Longrightarrow (i)$.

Now, we consider a $D-$optimal design ξ_D, and we show that $\boldsymbol{a_x}^T M(\xi_D)^- \boldsymbol{a_x} \leq m$ for every point $\boldsymbol{x} \in \mathcal{X}$, with equality when $\boldsymbol{x}$ is in the support of ξ_D. Note that a $D-$optimal design exists indeed, since we are maximizing a continuous function over a compact set. Moreover the optimal information matrix $M(\xi_D)$ is nonsingular, since there are m linearly independent vectors in the regression range (the matrix $M(\xi_D)$ must contain the columns of $\boldsymbol{I}$ in its range because we are interested in the whole parameter $\boldsymbol{\theta}$). Let $\boldsymbol{x} \in \mathcal{X}$, and consider the design $\xi_\alpha = (1-\alpha)\xi_D + \alpha\xi(\boldsymbol{x})$, where $\xi(\boldsymbol{x})$ is the design where all the experimental effort is concentrated at $\boldsymbol{x}$. The application $f : \alpha \to \log\det(M_\alpha)$, where $M_\alpha = (1-\alpha)M(\xi_D) + \alpha \boldsymbol{a_x} \boldsymbol{a_x}^T$ is the information matrix corresponding to the design ξ_α,

is well defined on $[0,1]$, and its derivative at $\alpha = 0$ exists and coincides with the directional derivative of $\log\det$ at $M(\xi_D)$ in the direction of $\boldsymbol{a_x}\boldsymbol{a_x}^T - M(\xi_D)$:

$$\left.\frac{df}{d\alpha}\right|_{\alpha=0} = \operatorname{trace} M(\xi_D)^{-1}(\boldsymbol{a_x}\boldsymbol{a_x}^T - M(\xi_D)) = \boldsymbol{a_x}^T M(\xi_D)^{-}\boldsymbol{a_x} - m.$$

The concavity of the $\log\det$ criterion and the optimality of the design ξ_D imply that f is nonincreasing on $[0,1]$, and so the latter derivative must be nonpositive. Hence,

$$\forall x \in \mathcal{X}, \boldsymbol{a_x}^T M(\xi_D)^{-}\boldsymbol{a_x} \leq m,$$

and we have proved the part $(ii) \Longrightarrow (iii)$. We further show that the latter inequality becomes an equality if $\boldsymbol{x}$ is a support point of ξ_D. We denote by $(\boldsymbol{x_i})_{i\in[s]}$ the support points of ξ_D and by $\boldsymbol{w}$ the vector of the associated weights, and we write:

$$\begin{aligned} m = \text{trace } I = \text{trace } M(\xi_D)M(\xi_D)^{-1} &= \operatorname{trace}(\sum_{i=1}^{s} w_i \boldsymbol{a_{x_i}}\boldsymbol{a_{x_i}}^T M(\xi_D)^{-1}) \\ &= \sum_{i|w_i>0} w_i \boldsymbol{a_{x_i}}^T M(\xi_D)^{-}\boldsymbol{a_{x_i}}. \end{aligned}$$

The latter expression is a weighted average of terms all smaller than m and takes the value m. Hence, $w_i > 0 \Rightarrow \boldsymbol{a_{x_i}}^T M(\xi_D)^{-}\boldsymbol{a_{x_i}} = m$.

Assume conversely that ξ is not $D-$optimal. If $M(\xi)$ is singular, then there is a regression vector $\boldsymbol{a_x}$ which is not in the range of $M(\xi)$, and so (iii) does not hold. If $M(\xi)$ has full rank, then in view of the strict concavity of the $\log\det$ function over $\mathbb{S}_m^+$, and similarly to the previous discussion, there exists a design ξ' such that $\log\det(M(\xi))$ has a positive derivative in the direction of $M(\xi') - M(\xi)$:

$$\operatorname{trace} M(\xi)^{-1}(M(\xi') - M(\xi)) = \operatorname{trace} M(\xi)^{-1}M(\xi') - m > 0.$$

Denoting the support points and the weights of ξ' by $\boldsymbol{x_i}'$ and w_i' respectively, we obtain:

$$\operatorname{trace} M(\xi)^{-1}M(\xi') = \sum_{i|w_i'>0} w_i' \boldsymbol{a_{x_i'}}^T M(\xi)^{-}\boldsymbol{a_{x_i'}} > m.$$

This expression is a weighted average strictly larger than m, which implies the existence of a support point $\boldsymbol{x'}$ of ξ' such that $\boldsymbol{a_{x_i'}}^T M(\xi)^{-}\boldsymbol{a_{x_i'}} > m$. Hence, (iii) does not hold and we have proved the part $(iii) \Longrightarrow (ii)$.

The existence of a $D-$optimal design, for which the Φ_G-criterion takes the value m, in conjunction with the fact that $\Phi_G(\xi) \geq m$ for all design ξ shows that a design ξ is $G-$optimal if and only if $\Phi_G(\xi) = m$. This proves the parts $(i) \Longrightarrow (iii)$ and the proof is complete. □

The previous result was extended to the case of multiresponse experiments by Fe-

dorov [Fed72]. The $G-$criterion for multiresponse experiments becomes

$$\Phi_G(\xi) = \max_{\boldsymbol{x}\in\mathcal{X}} \operatorname{trace} A(\boldsymbol{x})M(\xi)^- A(\boldsymbol{x})^T.$$

We omit the proof of this extended result, which is analogous to the previous one.

Theorem 2.4.3 (Fedorov [Fed72]). *Assume that the regression range $(A(\boldsymbol{x})^T\boldsymbol{z})_{x\in\mathcal{X},\ z\in\mathbb{R}^{l(\boldsymbol{x})}}$ contains at least m linearly independent vectors. Then the following statements are equivalent:*

(i) The design ξ is $G-$optimal;

(ii) The design ξ is $D-$optimal for the full parameter $\boldsymbol{\theta}$ (i.e. with $K = \boldsymbol{I}$);

(iii) For all $\boldsymbol{x}$ in $\mathcal{X}$, $\operatorname{trace} A(\boldsymbol{x})M(\xi)^- A(\boldsymbol{x})^T \leq m$.

Moreover, the bound provided by the inequality in (iii) is attained for the support points of the optimal design:

$$\boldsymbol{x_i} \in \operatorname{supp}(\xi) \Longrightarrow \operatorname{trace} A(\boldsymbol{x})M(\xi)^- A(\boldsymbol{x})^T \ = m.$$

This result was used by Fedorov to construct a sequential algorithm to build $D-$optimal designs: at each step, the point $\boldsymbol{x}$ which maximizes $\operatorname{trace} A(\boldsymbol{x})M(\xi)^- A(\boldsymbol{x})^T$ is sought, and the design ξ is replaced by a convex combination of ξ and the design $\xi(\boldsymbol{x})$ which concentrates all the experimental effort at $\boldsymbol{x}$ (cf. Section 3.1).

2.4.3 General Equivalence Theorem

In fact, the Kiefer-Wolfowitz theorem appears as a particular case of the General Equivalence Theorem proved by Kiefer in 1974 for some differentiable information criteria [Kie74], and extended by Pukelsheim in 1980 [Puk80] to any information criterion Φ that is non-negative, positively homogeneous and concave. The proof of Pukelsheim emphasizes on the convex duality flavour of the general equivalence theorem (his proof relies on Fenchel duality, and he proposed another approach based on subgradients with Titterington [PT83]). We give below a version of this theorem for the class of Kiefer's Φ_p criteria. For a proof, the reader is referred to Pukelsheim [Puk93].

Theorem 2.4.4 (General Equivalence Theorem [Kie74, Puk80]). *Consider a real number $p \in]-\infty, 1]$ (p finite). The design ξ is Φ_p-optimal for $K^T\boldsymbol{\theta}$ if and only if there is a generalized inverse G of $M(\xi)$ such that*

$$\forall \boldsymbol{x} \in \mathcal{X}, \qquad \operatorname{trace}\ A(\boldsymbol{x})GKQ_K(\xi)^{p+1}K^TGA(\boldsymbol{x})^T \leq \operatorname{trace}\ Q_K(\xi)^p.$$

In case of optimality, the latter inequality becomes an equality for any support point $\boldsymbol{x_i}$ of ξ.

Specifically, ξ is Φ_p-optimal for the whole parameter $\boldsymbol{\theta}$ if and only if

$$\forall \boldsymbol{x} \in \mathcal{X}, \qquad \operatorname{trace}\ A(\boldsymbol{x})M(\xi)^{p-1}A(\boldsymbol{x})^T \leq \operatorname{trace}\ M(\xi)^p.$$

We point out that there is a simpler version of this theorem when the information matrix $M(\xi)$ is assumed to be nonsingular at the optimum. The proof of this simplified version of the theorem is very close to that of the Kiefer-Wolfowitz equivalence theorem for $D-$optimality (Theorem 2.4.2). It relies on the directional derivative of $\Phi_p(Q_K(\xi))$ in the direction of the design $\xi(\boldsymbol{x})$ completely atomized at $\boldsymbol{x}$, which is well defined if $M(\xi)$ is invertible:

$$\begin{aligned}\phi'_{p,K}(\xi,\boldsymbol{x}) &= \lim_{\alpha\to 0^+} \frac{\Phi_p\left[Q_K\Big(\ (1-\alpha)\xi+\alpha\xi(\boldsymbol{x})\ \Big)\right]-\Phi_p[Q_K(\xi)]}{\alpha} \\ &= \text{trace}\ A(\boldsymbol{x})M(\xi)^{-1}KQ_K(\xi)^{p+1}K^TM(\xi)^{-1}A(\boldsymbol{x})^T-\text{trace}\ Q_K(\xi)^p.\end{aligned} \tag{2.22}$$

In the nonsingular case, we can formulate a general equivalence theorem that is very close to the original formulation of Kiefer and Wolfowitz for $D-$optimality:

Theorem 2.4.5 (General Equivalence theorem: Nonsigular case [Atw80])**.** *Let $p\in]-\infty,1]$ (p finite), and K an $r\times m$ matrix such that $K^T\boldsymbol{\theta}$ is estimable. Then, the following three statements are equivalent:*

(i) The design ξ^* is Φ_p-optimal for $K^T\boldsymbol{\theta}$;

(ii) $\phi'_{p,K}(\xi^*,\boldsymbol{x})\le 0$ for all $\boldsymbol{x}\in\mathcal{X}$;

(iii) ξ^* minimize $\max_{\boldsymbol{x}\in\mathcal{X}}\phi'_{p,K}(\xi^*,\boldsymbol{x})$ over $\Xi(K)$.

In addition, we have $\phi'_{p,K}(\xi^*,\boldsymbol{x})=0$ for all $\boldsymbol{x}\in\text{supp}\ \xi^*$.

This fundamental theorem, which gives an efficient method to check whether a given design is optimal, has several interesting consequences, which we next present.

Bound on D-optimal weights

We give below an interesting result of Pukelsheim [Puk80], which states that for single-response experiments, the weights of the $D-$optimal design for $K^T\boldsymbol{\theta}$ are bounded from above by $\frac{1}{r}$ (recall that r is the number of quantities that the experimenter wishes to estimate, i.e. r is the number of columns of K).

Theorem 2.4.6 (Bounds on $D-$optimal weights [Puk80])**.** *Every $D-$optimal design for $K^T\boldsymbol{\theta}$ has all its weights bounded from above by $\frac{1}{r}$. As a consequence, if the regression range $(\boldsymbol{a_x})_{x\in\mathcal{X}}$ consists in exactly r independent vectors which span the columns of K, then the $D-$optimal design for $K^T\boldsymbol{\theta}$ is unique and is defined by $w_i=\frac{1}{r}$ for all $i\in[r]$.*

Proof. Let ξ be a $D-$optimal design for $K^T\boldsymbol{\theta}$, and let $\boldsymbol{x_i}$ and w_i denote respectively the support points of ξ and their weights. By the generalized equivalence theorem 2.4.4 for $p=0$ ($D-$optimality), there exists a generalized inverse G of $M(\xi)$ such that:

$$\forall i\in[s],\qquad r=\text{trace}\ Q_K(\xi)^0=\boldsymbol{a_{x_i}}^TGKQ_K(\xi)K^TG\boldsymbol{a_{x_i}}=\boldsymbol{a_{x_i}}^TZ\boldsymbol{a_{x_i}}, \tag{2.23}$$

where we have set $Z = GKQ_K(\xi)K^TG$. In the latter expression, we can replace $Q_K(\xi)$ by $Q_K(\xi)K^TM(\xi)^-KQ_K(\xi)$, since the optimal $K-$information matrix must be invertible and $Q_K(\xi)^{-1} = K^TM(\xi)^-K$. Besides, notice that since G is a generalized inverse of $M(\xi)$, so is $GM(\xi)G$, and we can take this particular choice for $M(\xi)^-$:

$$r = \boldsymbol{a_{x_i}}^T GKQ_K(\xi)K^TGM(\xi)GKQ_K(\xi)K^TG\boldsymbol{a_{x_i}}.$$

We develop $M(\xi)$ as $\sum_{k\in[s]} w_k \boldsymbol{a_{x_k}}\boldsymbol{a_{x_k}}^T$ in order to obtain:

$$r = \sum_{k\in[s]} w_k(\boldsymbol{a_{x_i}}^T Z\boldsymbol{a_{x_k}})^2 \geq w_i(\boldsymbol{a_{x_i}}^T Z\boldsymbol{a_{x_i}})^2 = w_i r^2,$$

where we have used the expression of r that is given in (2.23). We finally obtain the desired upper bound:

$$w_i \leq \frac{r}{r^2} = \frac{1}{r}.$$

The second part of this theorem is a simple consequence of this upper bound. If $\mathcal{X} = [r]$ and the regression vectors are linearly independent and span the columns of K, then $K^T\boldsymbol{\theta}$ is estimable and the $D-$optimal design for $K^T\boldsymbol{\theta}$ affects a weight w_i no larger than $\frac{1}{r}$ to each of these r regression vectors. We can conclude that $w_i = \frac{1}{r}$ from the constraint $\sum_{i=1}^r w_i = 1$. □

An extension of this result to the framework of multiresponse experiments is possible. We made an announcement of the present result to the conference ISCO 2010 [BGS10] and it was discovered independently for the case $K = \boldsymbol{I}$ by Harman and Trnovská [HT09]. The proof mimics that of Theorem 2.4.6, and relies on an additional argument showing that when X is a positive semidefinite matrix, the ratio between $\mathrm{trace}\ X$ and $\mathrm{trace}\ X^2$ is bounded from below by a constant that depends on the rank of X. We will give a proof of this extension under a slightly different form in Chapter 7.

Theorem 2.4.7. *Let $\xi = \{\boldsymbol{x_k}, w_k\}$ be a $D-$optimal design for $K^T\boldsymbol{\theta}$. Then, the weight w_k of the experiment at $\boldsymbol{x_k}$ is bounded from above:*

$$w_k \leq \frac{\mathrm{rank}\ A(\boldsymbol{x_k})}{r}.$$

As a consequence, if (i) the regression region $\mathcal{X}$ is finite (of size s), (ii) $\sum_{k\in[s]} \mathrm{rank}\, A_k = r$, and (iii) the quantity $K^T\boldsymbol{\theta}$ is estimable, then the $D-$optimal design for $K^T\boldsymbol{\theta}$ is unique and is defined by

$$w_k = \frac{\mathrm{rank}\ A(\boldsymbol{x_k})}{r},\ \forall k \in [s].$$

A-Optimal weights on linearly independent regression vectors

Another interesting consequence of the general equivalence theorem was given by Pukelsheim and Torsney [PT91]. They showed that we can give the $A-$optimal design

for $K^T\boldsymbol{\theta}$ in close form when the regression range $\mathcal{X}$ is finite and the vectors $(\boldsymbol{a_x})_{x\in\mathcal{X}}$ are linearly independent. In this section, we associate $\mathcal{X}$ with $[s]$, so that the regression vectors are denoted by $\boldsymbol{a_1}, \ldots, \boldsymbol{a_s}$. We denote by $\mathcal{A}$ the aggregate of all row observation matrices:

$$\mathcal{A} = [\boldsymbol{a_1}, \ldots, \boldsymbol{a_s}]^T.$$

Note that this independence condition implies that the number s of vectors in $\mathcal{X}$ satisfies

$$r \leq s \leq m,$$

where the fist inequality is necessary because ξ bust be in the feasibility cone $\Xi(K)$, and the second inequality is enforced by the independence of the vectors $\boldsymbol{a_i}$. Besides, the design ξ is completely defined by the weight vector $\boldsymbol{w} \in \mathbb{R}^s$ since $\mathcal{X}$ is finite, so that we simply substitute $\boldsymbol{w}$ to ξ in the subsequent discussion.

The theorem of Pukelsheim and Torsney is actually proved in a more general context in [PT91], valid for any information criterion Φ that is nonnegative, positively homogeneous and concave, and establishes a nonlinear equation that the weights of the $\Phi-$optimal design must satisfy. A powerful corollary from their result is that this nonlinear equation can be solved in close form for the Kiefer's criterion Φ_{-1} ($A-$optimality). We give below an elementary proof of this powerful result.

Theorem 2.4.8 ($A-$optimal weights on independent regression vectors [PT91])**.** *If the regression vectors $\boldsymbol{a_1}, \ldots, \boldsymbol{a_s}$ are linearly independent and span the columns of K, then the the $A-$optimal design for $K^T\boldsymbol{\theta}$ is given in close form by*

$$\forall i \in [s], \qquad w_i = \frac{\sqrt{b_{ii}}}{\sum_{j=1}^{s}\sqrt{b_{jj}}},$$

where $b_{11}, \ldots, b_{ss}$ are the diagonal elements of the matrix

$$B = (\mathcal{A}\mathcal{A}^T)^{-1}\mathcal{A}KK^T\mathcal{A}^T(\mathcal{A}\mathcal{A}^T)^{-1},$$

which reduces to $B = (\mathcal{A}\mathcal{A}^T)^{-1}$ when the full parameter $\boldsymbol{\theta}$ is of interest ($K = \boldsymbol{I}$).

Proof. Let $\boldsymbol{w}$ be an $A-$optimal design for $K^T\boldsymbol{\theta}$. We first show that the statement of the theorem is true for all experiments which are in the support of the design $\boldsymbol{w}$, i.e. for all i such that $w_i > 0$. Let i denote the index of such an experiment. By the General equivalence theorem 2.4.4 for $p = -1$ ($A-$optimality), there exists a generalized inverse G of $M(\xi)$ such that:

$$\text{trace } Q_K(\boldsymbol{w})^{-1} = \boldsymbol{a_i}^T GKK^T G\boldsymbol{a_i}. \tag{2.24}$$

The columns of K are in the range of $M(\boldsymbol{w})$ because $\boldsymbol{w}$ must be in the feasibility cone $\Xi(K)$. Besides, $\boldsymbol{a_i}$ is in the range of $M(\boldsymbol{w}) = \sum_{i\in[s]} w_i\boldsymbol{a_i}\boldsymbol{a_i}^T$ because $w_i > 0$. Therefore, the vector $K^TG\boldsymbol{a_i}$ is invariant to the choice of the generalized inverse G of $M(\boldsymbol{w})$. Notice that $M(\boldsymbol{w})$ can be decomposed as $\mathcal{A}^T\,\text{Diag}(\boldsymbol{w})\mathcal{A}$. The linear independence of the vectors $\boldsymbol{a_1}, \ldots, \boldsymbol{a_s}$

implies that the matrix $\mathcal{A}\mathcal{A}^T$ is invertible, and so a particular choice for a generalized inverse of $M(\boldsymbol{w})$ is $G_0 = \mathcal{A}^T(\mathcal{A}\mathcal{A}^T)^{-1}\operatorname{Diag}(\boldsymbol{w})^\dagger(\mathcal{A}\mathcal{A}^T)^{-1}\mathcal{A}$. We use this particular choice for G in (2.24), and we use that $\boldsymbol{a_i} = \mathcal{A}^T\boldsymbol{e_i}$, where $\boldsymbol{e_i}$ is the i^{th} vector of the canonical basis of $\mathbb{R}^s$:

$$\operatorname{trace}\ Q_K(\boldsymbol{w})^{-1} = \boldsymbol{e_i}^T \operatorname{Diag}(\boldsymbol{w})^\dagger B \operatorname{Diag}(\boldsymbol{w})^\dagger \boldsymbol{e_i}.$$

In fact, the matrix $\operatorname{Diag}(\boldsymbol{w})^\dagger$ is the diagonal matrix where the k^{th} diagonal entry is either $\frac{1}{w_k}$ or 0 according as $w_k > 0$ or $w_k = 0$, so that the right hand side of the latter expression is equal to $b_{ii}w_i^{-2}$. We have thus shown that w_i is proportional to $\sqrt{b_{ii}}$.

It remains to show that the formula holds when $w_j = 0$, i.e. the j^{th} diagonal term of B is zero if $w_j = 0$. To see this, we assume without loss of generality that $w_1, \dots w_{s_0} > 0$ and $w_{s_0+1} = \ldots = w_s = 0$ for an index $s_0 \leq s$. Then, $\boldsymbol{a_1}, \dots, \boldsymbol{a_{s_0}}$ span the range of $M(\boldsymbol{w}) = \sum_{i=1}^{s_0} w_i \boldsymbol{a_i}\boldsymbol{a_i}^T$. Moreover, the columns of K are in the range of $M(\boldsymbol{w})$ by feasibility of the optimal vector $\boldsymbol{w}$, from which we deduce that there is a $s_0 \times r$ matrix H such that

$$K = \mathcal{A}^T \begin{pmatrix} H \\ 0 \end{pmatrix}.$$

Finally, for an index $j > s_0$ (i.e. such that $w_j = 0$), we obtain:

$$b_{jj} = \boldsymbol{e_j}^T B \boldsymbol{e_j} = \boldsymbol{e_j}^T(\mathcal{A}\mathcal{A}^T)^{-1}\mathcal{A}KK^T\mathcal{A}^T(\mathcal{A}\mathcal{A}^T)^{-1}\boldsymbol{e_j} = \boldsymbol{e_j}^T \begin{pmatrix} HH^T & 0 \\ 0 & 0 \end{pmatrix} \boldsymbol{e_j} = 0.$$

□

The latter result admits a straightforward generalization to the multiresponse case, which we do not think has been published elsewhere. The matrix $\mathcal{A}$ now stands for the aggregate observation matrix $[A_1^T, \dots, A_s^T]^T$.

Theorem 2.4.9 ($A-$optimal weights on indepedent observation matrices)**.** *If the rows of the observation matrices $A_1, \dots, A_s$ are linearly independent and span the columns of K, then the the $A-$optimal design for $K^T\boldsymbol{\theta}$ is given in close form by*

$$\forall i \in [s], \qquad w_i = \frac{\sqrt{\operatorname{trace} B_i}}{\sum_{j=1}^s \sqrt{\operatorname{trace} B_j}},$$

where $B_1, \dots, B_s$ are the diagonal blocks of size $l_1 \times l_1, \dots, l_s \times l_s$ of the matrix

$$B = (\mathcal{A}\mathcal{A}^T)^{-1}\mathcal{A}KK^T\mathcal{A}^T(\mathcal{A}\mathcal{A}^T)^{-1}.$$

c-Optimal weights on linearly independent regression vectors

As a corollary from the latter result, we obtain a simple closed-form formula for the weights of the $\boldsymbol{c}-$optimal design over independent regression vectors:

Corollary 2.4.10 ($\boldsymbol{c}-$optimal weights on independent regression vectors)**.** *If the regression vectors $\boldsymbol{a_1}, \ldots, \boldsymbol{a_s}$ are linearly independent and span the vector $\boldsymbol{c}$, then the the $\boldsymbol{c}-$optimal design is given in close form by*

$$\boldsymbol{w} = \frac{|(\mathcal{A}\mathcal{A}^T)^{-1}\mathcal{A}\boldsymbol{c}|}{\|(\mathcal{A}\mathcal{A}^T)^{-1}\mathcal{A}\boldsymbol{c}\|_1}.$$

(In the latter formula, the absolute value of the vector in the numerator is element-wise.) If in addition the number of regression vectors is $s = m$, then the matrix $\mathcal{A}$ is invertible and the latter formula simplifies to:

$$\boldsymbol{w} = \frac{|(\mathcal{A}^T)^{-1}\boldsymbol{c}|}{\|(\mathcal{A}^T)^{-1}\boldsymbol{c}\|_1}.$$

Proof. We know from Theorem 2.4.8 that the $\boldsymbol{c}-$optimal design $\boldsymbol{w}$ is proportional to the square root of the diagonal of

$$B = \Big((\mathcal{A}\mathcal{A}^T)^{-1}\mathcal{A}\boldsymbol{c}\Big)\Big((\mathcal{A}\mathcal{A}^T)^{-1}\mathcal{A}\boldsymbol{c}\Big)^T,$$

that is, $\boldsymbol{w} \propto |(\mathcal{A}\mathcal{A}^T)^{-1}\mathcal{A}\boldsymbol{c}|$. □

T-Optimal design for the full parameter $\boldsymbol{\theta}$

The next propositions show that the $T-$optimal design problem for the full parameter $\boldsymbol{\theta}$ is trivial. We start with the single-response case:

Proposition 2.4.11 ($T-$optimality for $\boldsymbol{\theta}$, single-response)**.** *A design is* formally $T-$*optimal if and only if all its support points correspond to regression vectors of maximal length, i.e.*

$$\forall i \in [s],\ w_i > 0 \Rightarrow \Big(\|\boldsymbol{a}_{\boldsymbol{x_i}}\| = \max_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{a}_{\boldsymbol{x}}\|\Big).$$

The extension to the multiresponse case is straightforward:

Proposition 2.4.12 ($T-$optimality for $\boldsymbol{\theta}$, multiresponse)**.** *A design is* formally $T-$*optimal if and only if all its support points correspond to observation matrices of maximal Frobenius norm, i.e.*

$$\forall i \in [s],\ w_i > 0 \Rightarrow \Big(\|A(\boldsymbol{x_i})\|_F = \max_{\boldsymbol{x}\in\mathcal{X}} \|A(\boldsymbol{x})\|_F\Big).$$

Proof. The (formal) $T-$optimal design problem for $\boldsymbol{\theta}$ can be formulated as:

$$\begin{aligned} \max_{\xi=\{\boldsymbol{x_k}, w_k\}} \quad & \text{trace} \sum_{i=1}^{s} w_i A(\boldsymbol{x_i}) A(\boldsymbol{x_i})^T \\ \text{s.t.} \quad & \sum_{i=1}^{s} w_i = 1; \quad \forall\ i \in [s],\ w_i \geq 0, \boldsymbol{x_i} \in \mathcal{X}. \end{aligned} \tag{2.25}$$

We have the following bound on the objective function:

$$\operatorname{trace} \sum_{i=1}^{s} w_i A(\boldsymbol{x_i}) A(\boldsymbol{x_i})^T = \sum_{i=1}^{s} w_i \|A(\boldsymbol{x_i})\|_F^2 \leq \underbrace{\sum_{i=1}^{s} w_i}_{1} \max_{\boldsymbol{x} \in \mathcal{X}} \|A(\boldsymbol{x})\|_F,$$

and it is clear that this bound is attained if and only if $\boldsymbol{w}$ assigns all its weight to points $\boldsymbol{x}$ where the observation matrix $A(\boldsymbol{x})$ is of maximal Frobenius norm. □

Chapter 3

Classic algorithms for computing optimal designs

When the regression region $\mathcal{X}$ is finite, or when the support points $\boldsymbol{x_1}, \ldots, \boldsymbol{x}_s$ are given, the optimal experimental design problem reduces to find the vector of weights $\boldsymbol{w}$. This arises in many practical situations, and in particular for the problem of optimal monitoring in networks that we present in the second part of this thesis. In the more general case where $\mathcal{X}$ is a compact region, many authors have proposed to solve a discretized version of the problem, by selecting a large (but finite) number of sample points in the regression region. A good motivation for this discretization is that the optimization problem is usually convex with respect to $\boldsymbol{w}$. Hence, if we ignore the optimization step over the support points, any local optimum is in fact a global optimum. This remarkable property is at the origin of several algorithms which converge to the optimal design vector $\boldsymbol{w}$. In this chapter, we study the Fedorov-Wynn exchange algorithm, a class of multiplicative algorithms, and the semidefinite programming (SDP) formulations for $E-, A-, D-$ and $T-$optimality.

In this chapter and the following ones, we associate the regression region $\mathcal{X}$ with $[s]$. Hence, every variable that was indexed by $\boldsymbol{x} \in \mathcal{X}$ will now be indexed by $i \in [s]$. Similarly, every variable depending on the design ξ will now be denoted as a function of $\boldsymbol{w}$. For example, the observation from the i^{th} experiment is

$$\boldsymbol{y_i} = A_i \boldsymbol{\theta} + \boldsymbol{\varepsilon_i},$$

and the information matrix reads

$$M(\boldsymbol{w}) = \sum_{i=1}^{s} w_i A_i^T A_i.$$

3.1 Federov-Wynn first order algorithm

Federov [Fed72] and Wynn [Wyn70] have described independently a method to compute $D-$optimal designs, inspired from the Kiefer-Wolfowitz theorem 2.4.2. The idea of this

algorithm is to start from an arbitrary design $\boldsymbol{w^{(0)}}$ and to move at each step in the direction of the design that is concentrated on the i^{th} experiment, where i is the index which maximizes $\operatorname{trace} A_i M(\boldsymbol{w})^{-} A_i^T$. More precisely, the following operation is performed at the k^{th} step of the algorithm:

$$\boldsymbol{w^{(k)}} = (1-\alpha_k)\boldsymbol{w^{(k-1)}} + \alpha_k \boldsymbol{e_i}, \quad \text{where} \quad i \in \operatorname*{argmax}_{i\in[s]} \operatorname{trace} A_i M(\boldsymbol{w})^{-} A_i^T.$$

In the latter expression $\boldsymbol{e_i}$ is the i^{th} standard unit vector of $\mathbb{R}^s$, and α_k is an appropriate sequence of step sizes. This algorithm was then generalized to a wider class of information functions Φ that are sufficiently regular by Atwood [Atw76, Atw80]. This includes the class of Φ_p-criteria for $K^T\boldsymbol{\theta}$, when the optimal design is such that $M(\boldsymbol{w})$ is non singular, and we restrict our discussion to this case.

This algorithm is in fact a feasible descent method: At each step, the design $\boldsymbol{w}$ is moved in the direction $\boldsymbol{w'} - \boldsymbol{w}$, where $\boldsymbol{w'}$ is a feasible design for which the directional derivative $\Phi'_{p,K}(\boldsymbol{w}, \boldsymbol{w'})$ is maximal ($\Phi'_{p,K}(\boldsymbol{w}, \boldsymbol{w'})$ denotes the directional derivative of $\Phi_p[Q_K(\boldsymbol{w})]$ at $\boldsymbol{w}$, in the direction of $\boldsymbol{w'} - \boldsymbol{w}$). By linearity of the derivative, we have:

$$\Phi'_{p,K}(\boldsymbol{w}, \boldsymbol{w'}) = \sum_i w'_i \phi'_{p,K}(\boldsymbol{w}, \boldsymbol{e_i}),$$

where $\phi'_{p,K}$ is the directional derivative in the direction of an atomic design, as defined in (2.22). Hence, a simple choice for $\boldsymbol{w'}$ is:

$$\boldsymbol{w'} = \operatorname*{arg\,max}_{\boldsymbol{v}|\sum_i v_i = 1} \quad \sum_i v_i \phi'_{p,K}(\boldsymbol{w}, \boldsymbol{e_i}) = \boldsymbol{e_j}, \quad \text{where} \quad j = \operatorname*{argmax}_{i\in[s]} \ \phi'_{p,K}(\boldsymbol{w}, \boldsymbol{e_i}).$$

The general Fedorov-Wynn algorithm follows. Its stopping criterion directly comes from the general equivalence theorem 2.4.5.

Algorithm 3.1.1 Fedorov-Wynn first order algorithm

Set a precision $\epsilon > 0$
Let $\boldsymbol{w^{(0)}}$ be a design such that $M(\boldsymbol{w^{(0)}})$ is nonsingular.
$k \leftarrow 0$
repeat
  $k \leftarrow k+1$
  Find $i_k = \operatorname{argmax}_{i\in[s]} \ \phi'_{p,K}(\boldsymbol{w^{(k)}}, \boldsymbol{e_i})$.
  Choose $\alpha_k \in [0,1]$ and construct $\boldsymbol{w^{(k)}} = (1-\alpha_k)\boldsymbol{w^{(k-1)}} + \alpha_k \boldsymbol{e_{i_k}}$.
until $\phi'_{p,K}(\boldsymbol{w^{(k)}}, \boldsymbol{e_i}) \leq \epsilon$

Classical stepsizes from literature on the the feasible direction methods can be used. Fedorov [Fed72] proposed the following rules:

(i) $\lim_{k\to\infty} \alpha_k = 0, \quad \sum_{k=1}^{\infty} \alpha_k = \infty;$

(ii) $\alpha_k = \operatorname{argmin}_{\alpha>0} \Phi_p(Q_K(\boldsymbol{w_\alpha^{(k)}})), \quad \text{where} \quad \boldsymbol{w_\alpha^{(k)}} = (1-\alpha)\boldsymbol{w^{(k-1)}} + \alpha \boldsymbol{e_{i_k}};$

(iii) $\alpha_k = \begin{cases} \alpha_{k-1}, \text{ if } \quad \Phi_p[Q_K(\boldsymbol{w_{\alpha_{k-1}}^{(k)}})] \geq \Phi_p[Q_K(\boldsymbol{w^{(k-1)}})] \\ \alpha_{k-1}/\gamma, \ \gamma > 1 \text{ otherwise.} \end{cases}$

The convergence of this algorithm is proved in [WW78] for the rules (i) and (ii). Another proof for the rule (ii) is presented in [Atw80]. Cook and Fedorov claim [CF95] that the convergence of the algorithm for the three above rules on the step sizes is standard, without giving a proof. We also indicate that Richtarik [Ric08] recently proposed a Fedorov-Wynn–type algorithm with specified steplengths α_k, for which it is guaranteed that a $\delta-$approximate solution is returned after $O(1/\delta)$ iterations.

An important property of the Fedorov-Wynn–type algorithms is the following. By rewriting the update rule of $\boldsymbol{w}$ as:

$$\boldsymbol{w}^{(\boldsymbol{k})} = (1-\alpha_k)\Big(\boldsymbol{w}^{(\boldsymbol{k-1})} + \frac{\alpha_k}{1-\alpha_k}\boldsymbol{e}_{\boldsymbol{i_k}}\Big),$$

we see that the information matrix $M(\boldsymbol{w}^{(\boldsymbol{k})})$ can be written as:

$$M(\boldsymbol{w}^{(\boldsymbol{k})}) = (1-\alpha_k)\Big(M(\boldsymbol{w}^{(\boldsymbol{k-1})}) + \frac{\alpha_k}{1-\alpha_k}A_{i_k}A_{i_k}^T\Big).$$

We usually have $l_i << m$, and so the latter formula is a *low-rank* update of the information matrix. Therefore, much computational saving can be obtained by using the Sherman-Morrison formula to update the inverse of $M(\boldsymbol{w})$, which is often required to evaluate the $\phi'_{p,K}(\boldsymbol{w}^{(\boldsymbol{k})}, \boldsymbol{e_i})$. In some situations, it can be sufficient to compute low rank updates of the LU decomposition of $M(\boldsymbol{w})$.

We point out that for the sequence of step sizes $\alpha_k = (1+k)^{-1}$ (which satisfies the rule (i)), the algorithm can be interpreted as a sequential algorithm for constructing *non-normalized* designs: At each step of the algorithm, a new measurement is added on the experiment which maximizes the directional derivative $\phi'_{p,K}(\boldsymbol{w}^{(\boldsymbol{k})}, \boldsymbol{e_i})$. The step sizes $\alpha_k = (1+k)^{-1}$ mimics this sequential procedure while keeping the designs normalized (i.e. $\sum_i \boldsymbol{w}^{(\boldsymbol{k})} = 1$). This was proposed by Fedorov [Fed72] for the construction of $D-$optimal designs. A refinement of this sequential procedure is possible: at each step, the experimenter has both the possibility to add a "good" measurement point (corresponding to the largest value of the derivative) and to remove a "bad" one (corresponding to the small value of the derivative). This procedure is known as the *Fedorov Exchange algorithm*. One can further define forward and backward excursions, where n^+ new measurement points are added and n^- are deleted, as in Mitchell [Mit74].

3.2 Multiplicative weight updates

Multiplicative algorithms were proposed in 1976 by Titterigton to compute the weights of the $D-$optimal design [Tit76] (for the full vector $\boldsymbol{\theta}$). The idea is to multiply, at each step, every coordinate w_i of the current design $\boldsymbol{w}^{(\boldsymbol{t})}$ by a factor which is proportional to the derivative

$$\left.\frac{\partial \log\det M(\boldsymbol{w})}{w_i}\right|_{\boldsymbol{w}=\boldsymbol{w}^{(\boldsymbol{t})}} = \operatorname{trace} A_i M(\boldsymbol{w}^{(\boldsymbol{t})})^{-1} A_i^T.$$

At each step, the normalization factor is simply

$$\sum_{i=1}^{s} w_i^{(t)} \operatorname{trace} A_i M(\boldsymbol{w}^{(t)})^{-1} A_i^T = \operatorname{trace}\Big(M(\boldsymbol{w}^{(t)})^{-} M(\boldsymbol{w}^{(t)})\Big) = \operatorname{trace} \boldsymbol{I} = m,$$

such that the iterations are:

$$\forall i \in [s], \quad w_i^{(t+1)} = w_i^{(t)} \frac{\operatorname{trace} A_i M(\boldsymbol{w}^{(t)})^{-1} A_i^T}{m}. \tag{3.1}$$

Titterington proved in [Tit76] that the sequence of determinants $\det M(\boldsymbol{w}^{(t)})$ generated by this sequence is nondecreasing, and converges to the optimal value of the $D-$criterion. He also proposed [Tit78] a variant of the form:

$$\forall i \in [s], \quad w_i^{(t+1)} = w_i^{(t)} \frac{\operatorname{trace} A_i M(\boldsymbol{w}^{(t)})^{-1} A_i^T - \beta}{m - \beta}, \tag{3.2}$$

which is faster than the iterations (3.1) in practice, and conjectured the monotonic behaviour of the sequence of determinants for $\beta = 1$. Under a slightly different setting, Dette, Pepelyshev and Zhigljavsky [DPZ08] proved the monotonicity of $\det M(\boldsymbol{w}^{(t)})$, for iterations of the form (3.2), with a dynamic parameter $\beta^{(t)}$ instead of β. The conjecture was finally resolved in 2010 by Yu [Yu10b].

A general class of multiplicative algorithms was proposed in 1978 by Silvey, Titterington and Torsney [STT78], for the $\Phi-$optimal design problem:

$$\forall i \in [s], \quad w_i^{(t+1)} = w_i^{(t)} \frac{d_i(\boldsymbol{w}^{(t)})^{\lambda}}{\sum_{j \in [s]} w_j^{(t)} d_j(\boldsymbol{w}^{(t)})^{\lambda}}, \tag{3.3}$$

where $d_i(\boldsymbol{w}^{(t)}) = \frac{\partial \Phi[M(\boldsymbol{w})]}{w_i}\Big|_{\boldsymbol{w}=\boldsymbol{w}^{(t)}}$ and λ is a power parameter in $]0, 1]$. For the $A-$optimal design problem, the monotonicity of the sequence $\Phi_A[M(\boldsymbol{w}^{(t)})]$ was proved by Torsney [Tor83] for the power parameter $\lambda = 1/2$. Yu proved recently [Yu10a] the convergence of this general class of multiplicative algorithms for the design criteria $\Phi[M(\boldsymbol{w})]$ such that $M \mapsto -\Phi(M^{-1})$ is concave (with respect to Löwner ordering). This includes as a special case the Φ_p-optimal design problem for $K^T\boldsymbol{\theta}$, when $p \in [-1, 0]$ (in particular, for $A-$ and $D-$optimality).

The different versions of the multiplicative weight updates are presented in a unified way in Algorithm 3.2.1, for the Φ_p-optimal design problem for $K^T\boldsymbol{\theta}$. The stopping criterion is based on the general equivalence theorem 2.4.4, and we have used the fact that for every design $\boldsymbol{w}$, we have $\sum_{i \in [s]} w_i^{(t)} d_i(\boldsymbol{w}^{(t)}) = \operatorname{trace} Q_K(\boldsymbol{w}^{(t)})^p$.

We also point out that for $p = -1$ ($A-$optimality), the derivative of the criterion $\Phi_A[Q_K\Big(M(\boldsymbol{w})\Big)]$ takes the simple form

$$d_i(\boldsymbol{w}^{(t)}) = \|A_i M(\boldsymbol{w}^{(t)})^{-} K\|_F^2.$$

In particular, for $\boldsymbol{c}-$optimality, we obtain:

$$d_i(\boldsymbol{w^{(t)}}) = \|A_i M(\boldsymbol{w^{(t)}})^{-}\boldsymbol{c}\|^2.$$

For the case of $E-$optimality, the criterion is not differentiable in general, but a subgradient is given by

$$d_i(\boldsymbol{w^{(t)}}) = \|A_i M(\boldsymbol{w^{(t)}})^{-}K\boldsymbol{v}\|^2,$$

where $\boldsymbol{v}$ is an eigenvector associated to the largest eigenvalue of $K^T M(\boldsymbol{w^{(t)}})^{-}K$.

Algorithm 3.2.1 Titterington-type multiplicative algorithm

Set a precision $\epsilon > 0$
Choose a power parameter λ
Let $\boldsymbol{w^{(0)}}$ be a design such that $M(\boldsymbol{w^{(0)}})$ is nonsingular.
$t \leftarrow 0$
repeat
  $t \leftarrow t + 1$
  for $i \in [s]$ **do**
    $d_i^{(t)} \leftarrow \text{trace } A_i M(\boldsymbol{w^{(t)}})^{-1} K Q_K(\boldsymbol{w^{(t)}})^{p+1} K^T M(\boldsymbol{w^{(t)}})^{-1} A_i^T$
  end for
  for $i \in [s]$ **do**
    Choose an acceleration parameter $\beta^{(t)}$.
    $w_i^{(t+1)} = w_i^{(t)} \dfrac{\left(d_i^{(t)}\right)^{\lambda} - \beta^{(t)}}{\sum_{j\in[s]} w_j^{(t)} \left(d_j^{(t)}\right)^{\lambda} - \beta^{(t)}}$
  end for
until $\max_{i\in[s]} d_i^{(t)} \leq (1+\epsilon) \sum_{i\in[s]} w_i^{(t)} d_i^{(t)}$

3.3 Mathematical programming approaches

In this section, we review the linear programming (LP), semidefinite programming (SDP), and determinant maximization (MAXDET) formulations that have been proposed to solve some optimal experimental design problems.

When Pukelsheim have proved the general equivalence theorem 2.4.4 for any information function that is nonnegative, positively homogeneous and concave, he incidentally gave a dual formulation of the $E-$optimal design which is nothing but a semidefinite program [Puk80]. However, this feature does not seem to have been noticed at this period, probably because the semidefinite programming theory and algorithms were still at a very early stage of their development. The SDP approach to optimal experimental design was then "rediscovered" by Vandenberghe, Boyd and Wu in 1999 [VBW98], who were able to formulate semidefinite programs for the $E-$ and $A-$optimal design problems, and a MAXDET problem for the $D-$optimal design (for the full parameter $\boldsymbol{\theta}$). A review of these formulations is presented by Fedorov and Lee [FL00]; another one is available in Chapter 7.5 of Boyd and Vandenberghe [BV04].

Recently, Harman and Jurík [HJ08] showed that the Elfving theorem 2.4.1 yields a linear programming formulation of the $\boldsymbol{c}-$optimal design. On the other hand, the $\boldsymbol{c}-$optimal design problem also admits a semidefinite programming formulation which was studied by Qi [Qi09]. In the analysis of his multiplicative-low rank update algorithm, Richtarik [Ric08] pointed out the equivalence between the latter LP and SDP approaches, and noticed that a rank 1 solution of the SDP always exist. We will extend this result of existence of low rank solutions to a wider class of semidefinite programs in Chapter 4.

3.3.1 E-optimality

The $E-$ optimal design for the full parameter $\boldsymbol{\theta}$ aims at maximizing the smallest eigenvalue of the information matrix $M(\boldsymbol{w})$. We will make use of the characterization of the smallest eigenvalue of a symmetric matrix by Rayleigh-Ritz quotients: $M \in \mathbb{S}_m$:

$$\lambda_{\min}(M) = \min_{\boldsymbol{v}\in\mathbb{R}^m,\ \boldsymbol{v}\neq\mathbf{0}} \quad \frac{\boldsymbol{v}^T M \boldsymbol{v}}{\boldsymbol{v}^T \boldsymbol{v}}.$$

The latter expression implies that for every scalar $t \leq \lambda_{\min}(M)$ and for all vector $\boldsymbol{v} \in \mathbb{R}^m$,

$$\boldsymbol{v}^T M \boldsymbol{v} \geq t\ \boldsymbol{v}^{\boldsymbol{T}} \boldsymbol{v}.$$

This can be rewritten as $\forall \boldsymbol{v},\ \boldsymbol{v}^T(M - t\boldsymbol{I})\boldsymbol{v} \geq 0$, or equivalently: $M \succeq t\boldsymbol{I}$. Similarly, if $t > \lambda_{\min}(M)$, there must exist a vector $\boldsymbol{v_0}$ such that $\boldsymbol{v_0}^T(M - t\boldsymbol{I})\boldsymbol{v_0} < 0$, and $M \not\succeq t\boldsymbol{I}$. This proves:

$$\forall M \in \mathbb{S}_m, \qquad \lambda_{\min}(M) = \max_{t\in\mathbb{R}} \quad t \qquad (3.4)$$
$$\text{s.t.} \quad M \succeq t\boldsymbol{I}.$$

Thanks to this SDP formulation of the smallest eigenvalue of a symmetric matrix, and by associativity of the $\max$ operator, we can formulate the $E-$optimal design problem (2.17) as:

$$\max_{t,\boldsymbol{w}} \quad t \qquad (3.5)$$
$$\text{s.t.} \quad M(\boldsymbol{w}) \succeq t\boldsymbol{I}$$
$$\sum_{i=1}^{s} w_i = 1, \quad \forall\ i \in [s], w_i \geq 0.$$

In fact, the more general $E-$optimal design problem for the estimation of $K^T\theta$ can also be expressed as a semidefinite program, by substituting KK^T to $\boldsymbol{I}$ in the right hand side of

the linear matrix inequality of Problem (3.5):

$$\begin{aligned} \max_{t,\boldsymbol{w}} \quad & t \\ \text{s.t.} \quad & M(\boldsymbol{w}) \succeq tKK^T \\ & \sum_{i=1}^{s} w_i = 1, \quad \forall\ i \in [s], w_i \geq 0. \end{aligned} \tag{3.6}$$

We recall that the optimal design $\boldsymbol{w}$ must lie in the feasibility cone $\Xi(K)$, which means that the range of K must be included in that of $M(\boldsymbol{w})$. This, of course, is automatically implied by the linear matrix inequality $M(\boldsymbol{w}) \succeq tKK^T$ of Problem (3.6), in accordance with our discussion following Equation (2.15).

We show below that the Lagrangian dual of the $E-$optimality SDP (3.6) already appeared in Pukelsheim [Puk80], as a special case of his duality theorem. For an information function Φ that is nonnegative on $\mathbb{S}_m^+$, positive on $\mathbb{S}_m^{++}$, positively homogeneous and concave, its *polar* function is defined as:

$$\Phi^*(X) = \inf_{Z \succ 0} \ \frac{\langle Z, X \rangle}{\Phi(Z)}.$$

We give below a version of Pukelsheim's duality theorem for the case in which $\mathcal{X}$ is finite:

Theorem 3.3.1 (Duality theorem [Puk80])**.**

$$\begin{array}{lcl} \sup & \Phi(Q_K(\boldsymbol{w})) & = \quad \inf_{X \succeq 0} \quad 1/\Phi^*(K^T X K) \\ \text{s. t.} & \boldsymbol{w} \in \Xi(K) & \text{s. t.} \quad \langle A_i^T A_i, X \rangle \leq 1 \ (\forall i \in [s]). \end{array}$$

Now, for $\Phi = \Phi_E = \lambda_{\min}(\cdot)$, it is easy to see that $\Phi_E^*(X) = \text{trace } X$, and the expression at the right hand side of the equality sign in Theorem 3.3.1 is the inverse of

$$\begin{aligned} \max_{X \succeq 0} \quad & \langle KK^T, X \rangle \\ \text{s. t.} \quad & \langle A_i^T A_i, X \rangle \leq 1 \ (\forall i \in [s]), \end{aligned} \tag{3.7}$$

which is a semidefinite program. Its dual is:

$$\begin{aligned} \min_{\boldsymbol{\mu} \geq 0} \quad & \sum_{i=1}^{s} \mu_i \\ \text{s.t.} \quad & \sum_{i=1}^{s} \mu_i A_i^T A_i \succeq KK^T. \end{aligned} \tag{3.8}$$

The Slater condition holds for the pair of problems (3.7) and (3.8), because they are both strictly feasible (under the assumption that $K^T\boldsymbol{\theta}$ is estimable). This means that strong duality holds, and these programs share the same optimal value. Finally, we can see that the inverse of the optimal value of Problem (3.8) coincides with the optimum of Problem (3.6),

thanks to the normalization $t = \frac{1}{\sum_i \mu_i}$, $\boldsymbol{w} = t\boldsymbol{\mu}$.

3.3.2 D-optimality

A $D-$ optimal design for the full parameter $\boldsymbol{\theta}$ maximizes the determinant of the information matrix $M(\boldsymbol{w})$. The problem of maximizing a determinant under some linear matrix inequality (LMI) constraints has been studied by Vandenberghe, Boyd and Wu [VBW98]. They showed that this class of problems can be considered as a generalization of semidefinite programs and give an interior point algorithm for their resolution. The MAXDET formulation of the $D-$optimal design (for the full parameter $\boldsymbol{\theta}$) is:

$$\begin{aligned} \max_{\boldsymbol{w}} \quad & \log\det M(\boldsymbol{w}) \\ & \sum_{i=1}^{s} w_i = 1, \quad \forall\ i \in [s], w_i \geq 0, \end{aligned} \tag{3.9}$$

where the logarithm in the objective function ensures the convexity of the criterion. The dual of this problem is of particular interest:

$$\begin{aligned} \max_{W \succeq 0} \quad & \log\det W \\ & \langle A_i^T A_i, W\rangle \leq m, \quad \forall\ i \in [s], w_i \geq 0. \end{aligned} \tag{3.10}$$

Under the generic assumption that the full vector $\boldsymbol{\theta}$ is estimable, i.e. that there is a design $\boldsymbol{w}$ such that $M(\boldsymbol{w})$ has full rank, strong duality holds between Problems (3.9) and (3.10) (Slater's condition is fulfilled), and the complementary slackness relation yields:

$$w_i(\langle A_i^T A_i, W\rangle - m) = 0.$$

In the single-response case ($A_i = \boldsymbol{a_i}^T$), the dual problem (3.10) can be interpreted as finding the minimal-volume ellipsoid centered at the origin which contains the points $\boldsymbol{a_1}, \ldots, \boldsymbol{a_s}$. The complementary slackness relation further indicates that the support of the $D-$optimal design consists in experiments whose regression vector lies on the surface of this minimal ellipsoid (cf. Figure 3.1).

3.3.3 A-optimality

An $A-$optimal design problem for $K^T\boldsymbol{\theta}$ minimizes

$$\boldsymbol{c_1}^T M(\boldsymbol{w})^- \boldsymbol{c_1} + \ldots + \boldsymbol{c_r}^T M(\boldsymbol{w})^- \boldsymbol{c_r},$$

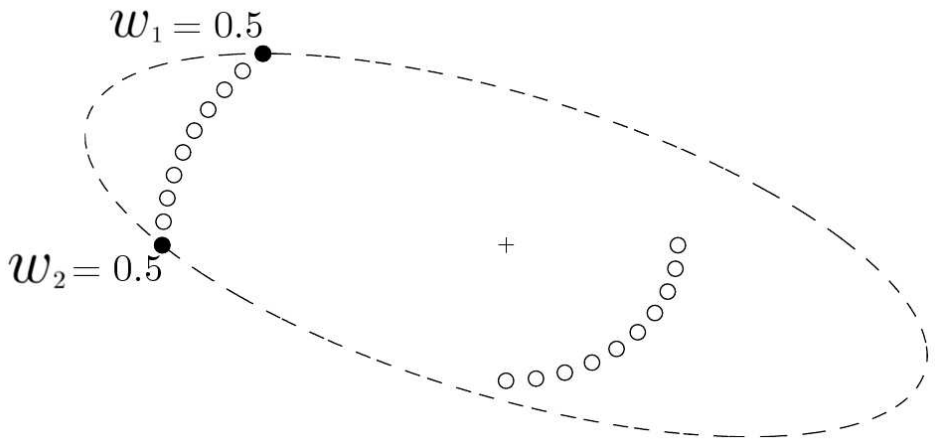


Figure 3.1: This figure is extracted from [BV04]. It shows the geometric interpretation of the $D-$optimal design for single-response experiments. The origin is marked with a cross and the regression vectors $\boldsymbol{a_1}, \ldots, \boldsymbol{a_s}$ are indicated with circles. The $D-$optimal design uses the two measurement vectors indicated with solid circles. Since the corresponding regression vectors are linearly independent, it follows from Theorem 2.4.6 that the $D-$optimal design puts equal weights $w_1 = w_2 = 0.5$ on each of them. The ellipse corresponds to the minimal-volume ellipsoid centered at the origin and that contains all the measurement points.

where $\boldsymbol{c_1}, \ldots, \boldsymbol{c_r}$ denote the columns of K. Each term of this sum can be bounded from above thanks to a linear matrix inequality, by using the Schur complement lemma:

$$t_i \geq \boldsymbol{c_1}^T M(\boldsymbol{w})^- \boldsymbol{c_1} \iff \left(\begin{array}{c|c} M(\boldsymbol{w}) & \boldsymbol{c_i} \\ \hline \boldsymbol{c_i}^T & t_i \end{array}\right) \succeq 0.$$

This property allows one to formulate the $A-$optimal design problem (2.18) as a semidefinite program:

$$\begin{aligned} \min_{\boldsymbol{w},\boldsymbol{t}} \quad & \sum_{j=1}^{r} t_j \\ \text{s.t.} \quad & \left(\begin{array}{c|c} M(\boldsymbol{w}) & \boldsymbol{c_j} \\ \hline \boldsymbol{c_j}^T & t_j \end{array}\right) \succeq 0, \quad j \in [r], \\ & \sum_{i=1}^{s} w_i = 1, \quad \forall\, i \in [s], w_i \geq 0. \end{aligned} \tag{3.11}$$

This was first noticed by the authors of [VBW98] in the case where the full parameter $\boldsymbol{\theta}$ is of interest, i.e. $K = \boldsymbol{I}$, $r = m$, and $\boldsymbol{c_i} = \boldsymbol{e_i}$ (the i^{th} standard unit vector of $\mathbb{R}^m$). An alternative formulation involving an auxiliary matrix variable Y, but only one LMI was proposed by Fedorov and Lee [FL00]. We extend their formulation to the case in which

$K^T\boldsymbol{\theta}$ is of interest:

$$\begin{aligned} \min_{\boldsymbol{w},\ Y\in\mathbb{S}_r} \quad & \operatorname{trace}\ Y \\ \text{s.t.} \quad & \left(\begin{array}{c|c} M(\boldsymbol{w}) & K \\ \hline K^T & Y \end{array}\right) \succeq 0, \\ & \sum_{i=1}^{s} w_i = 1, \quad \forall\ i \in [s], w_i \geq 0. \end{aligned} \tag{3.12}$$

We point out that the formulation proposed by the authors of [AS08] turns out to be valid only if every information matrix $M_i = A_i^T A_i$ is diagonal: their SDP is analogous to Problem (3.12), but Y is forced to be a diagonal matrix ($Y = \operatorname{Diag}(\boldsymbol{y})$). Contrarily to what they claim, this does not yield $A-$optimal designs: the positivity of the Schur complement $\operatorname{Diag}(\boldsymbol{y}) \succeq K^T M(\boldsymbol{w})^- K$ implies $\operatorname{trace}(\operatorname{Diag}(\boldsymbol{y})) \geq \operatorname{trace} K^T M(\boldsymbol{w})^- K$, but there are simple examples where this inequality is strict at the optimum.

3.3.4 c-optimality

Single-response case: LP approaches

In presence of scalar observations ($A_i = \boldsymbol{a_i}^T$), Elfving's geometric characterization of $\boldsymbol{c}-$optimality (Theorem 2.4.1) yields a linear program. Finding the optimum indeed reduces to computing the intersection of the vectorial straight line directed by $\boldsymbol{c}$ and the boundary of the polyhedron with vertices $\pm\boldsymbol{a_i}$ (see Figure 2.2):

$$\begin{aligned} \max_{\boldsymbol{\lambda},t} \quad & t \\ \text{s.t.} \quad & t\boldsymbol{c} = \sum_k \boldsymbol{a_i}\lambda_i \\ & \sum_k \underbrace{|\lambda_k|}_{w_k} \leq 1. \end{aligned} \tag{3.13}$$

Elfving's Theorem further indicates that the optimal value of the criterion $\boldsymbol{c}^T M(\boldsymbol{w})^- \boldsymbol{c}$ is t^{-2}. After the change of variable $\tau = \frac{1}{t}$, $\boldsymbol{\mu} = \tau\boldsymbol{\lambda}$, the dual of this problem is:

$$\begin{aligned} \max_{\boldsymbol{x}\in\mathbb{R}^m} \quad & \boldsymbol{c}^T\boldsymbol{x} \\ \text{s.t.} \quad & |\boldsymbol{a_i}^T\boldsymbol{x}| \leq 1,\ i \in [s]. \end{aligned} \tag{3.14}$$

General case: SDP approaches

A $\boldsymbol{c}$-optimal design is a particular case of a $E-$ (or $A-$) optimal design, for $r = 1$. Hence, we obtain the following formulations for the $\boldsymbol{c}-$optimal design problem: from the

$E-$optimality SDP (3.6) we get

$$\begin{aligned} \max_{t,\boldsymbol{w}} \quad & t \\ \text{s.t.} \quad & M(\boldsymbol{w}) \succeq t\boldsymbol{c}\boldsymbol{c}^T \\ & \sum_{i=1}^{s} w_i = 1, \quad \forall\ i \in [s], w_i \geq 0. \end{aligned} \tag{3.15}$$

We also obtain its dual in the form of Problem (3.7):

$$\begin{aligned} \max_{X \succeq 0} \quad & \boldsymbol{c}^T X \boldsymbol{c} \\ \text{s. t.} \quad & \langle A_i^T A_i, X \rangle \leq 1 \ (\forall i \in [s]). \end{aligned} \tag{3.16}$$

The $A-$optimality SDP (3.11) yields an alternative formulation:

$$\begin{aligned} \min_{\boldsymbol{w},\ \tau \in \mathbb{R}} \quad & \tau \\ \text{s. t.} \quad & \left(\begin{array}{c|c} M(\boldsymbol{w}) & \boldsymbol{c} \\ \hline \boldsymbol{c} & \tau \end{array} \right) \succeq 0, \\ & \sum_{i=1}^{s} w_i = 1, \quad \forall\ i \in [s], w_i \geq 0, \end{aligned} \tag{3.17}$$

which reduces to Problem (3.15) after the change of variable $t = \frac{1}{\tau}$ and the reformulation of the LMI by the Schur complement lemma.

Richtarik [Ric08] and Qi [Qi09] noticed independently that the Elfving theorem (in the single-response setting) implies that a solution of rank one of Problem (3.16) always exists. The search for a solution of the form $X = \boldsymbol{x}\boldsymbol{x}^T$ indeed reduces to Problem (3.14) (up to a square in the objective function which can be removed, since, if $\boldsymbol{x}$ is a solution of Problem (3.14), so is $-\boldsymbol{x}$). We will see in Chapter 5 that this property is also valid in the general *multiresponse* case. An important consequence is that the semidefinite program (3.16) reduces to a Second order cone program (SOCP), which we study in Chapter 5. This contradicts Qi's claim [Qi09], according to which computing the best rank-one solution of Problem (3.16) is a nonconvex problem which is extremely difficult to solve.

3.3.5 Flexibility of mathematical programming approaches

In general, the mathematical programming approaches studied in this section are slower than the specialized algorithms of Section 3.2 (a comparative study of the algorithms is done in Chapter 6). However, we point out that the SOCP approaches, which will be studied in Chapter 5, are competitive with the specialized algorithms in many situations. But the great advantage of mathematical programming formulations resides mostly in their flexibility, and

the possibility to add "without effort" new constraints in the problem. We now give a few examples of these possibilities.

Multiple resource constraints

Elfving studied the case in which the available experiments have different costs [Elf52]. If the cost of the i^{th} experiment is p_i, and the experimenter disposes of a budget b, the constraint becomes:

$$\sum_{i=1}^{s} w_i p_i \leq b.$$

Now, w_i can not be interpreted as the *percentage of experimental effort to spend on the* i^{th} *experiment* anymore. Instead, the quantity $w_i \frac{p_i}{b}$ should be seen as the percentage of budget to allocate to the experiment i. Elfving noticed that the change of variable $w_i' = w_i \frac{p_i}{b}$ brings the problem back to the previous situation, and is equivalent to a scaling of the observation equations (2.3).

Consider now the more general case in which $\boldsymbol{w}$ is a control variable for the experiments, such that the information matrix takes the standard form $M(\boldsymbol{w}) = \sum_{i=1}^{s} w_i A_i^T A_i$ for some observation matrices A_i. We assume that $\boldsymbol{w}$ is constrained by several linear inequalities

$$R\boldsymbol{w} \leq \boldsymbol{b}, \tag{3.18}$$

where $\boldsymbol{b} \in \mathbb{R}^n$, R is a $n \times s$ matrix and the inequality is elementwise. Contrarily to the previous situation with a single *budget constraint*, there is no simple change of variable which brings the problem back to the standard case ($\sum_i w_i = 1$), because we do not know which inequalities will be saturated in (3.18) at optimality. This constrained problem has been studied by Cook and Fedorov [CF95], who proposed an extension of the Fedorov exchange algorithm (cf. Section 3.1). However, this algorithm exhibits a very slow convergence in practice.

This constrained framework arises in the problem of optimally setting the sampling rates of a measuring device on a network (see Part II): here, $\boldsymbol{w}$ is the vector of the sampling rates of the monitoring tool at different locations of the network, and the constraint $R\boldsymbol{w} \leq \boldsymbol{b}$ reflects the fact that only a certain number of packets should be sampled at each router. This *multiple resource constraint* can be added in any of the previous SDPs without any reformulation effort. For example, Singhal and Michailidis [SM08] considered the following resource constrained SDP for $A-$optimality:

$$\begin{aligned} \min_{\boldsymbol{w},\boldsymbol{t}} \quad & \sum_{j=1}^{r} t_j \\ \text{s. t.} \quad & \left(\begin{array}{c|c} M(\boldsymbol{w}) & \boldsymbol{c_j} \\ \hline \boldsymbol{c_j}^T & t_j \end{array}\right) \succeq 0, \quad j \in [r], \\ & R\boldsymbol{w} \leq \boldsymbol{b}, \quad \forall\, i \in [s], w_i \geq 0. \end{aligned} \tag{3.19}$$

Bounding the eigenvalues

Harman, Jurík and Trnovská [HJT07] have proposed to add a lower bound on the minimum eigenvalue of the information matrix ($\lambda_{\min}(Q_K(\boldsymbol{w})) \geq \lambda_0$). Geometrically, this is equivalent to impose an upper bound on the diameter of the confidence ellipsoids (2.11), or to guarantee that the $E-$criterion is at least λ_0. This constraint guards us against the case in which one of the quantities $\zeta_i = \boldsymbol{c_i}^T$ is badly estimated. It is of particular interest for the $D-$optimal design problem, where the confidence ellipsoids are of minimal volume at the optimum, but can theoretically have an arbitrarily large diameter. In practice, a way to introduce this constraint is to impose the LMI

$$M(\boldsymbol{w}) \succeq \lambda_0 K K^T$$

on the design (see Section 3.3.1).

Avoiding "concentrated designs"

Vandenberghe, Boyd and Wu [VBW98] have described another useful constraint that can be imposed on the model: The goal is to avoid a large fraction of the experimental effort, say 90%, of being concentrated on a small number of experiments, say 10% of the possible observations. This "90-10" constraint has the effect to spread out the measurements over the possible experiments:

$$\sum_{i=1}^{\lfloor s/10 \rfloor} w_{[i]} \leq 0.9,$$

where $w_{[i]}$ is the i^{th} largest component of $\boldsymbol{w}$. The authors of [VBW98] show that this constraint is satisfied if and only if there exists a vector $\boldsymbol{x} \in \mathbb{R}^s$ and a scalar t such that:

$$\begin{aligned} \left\lfloor \frac{s}{10} \right\rfloor t + \sum_{i=1}^{s} x_i &\leq 0.9, \\ t + x_i &\geq w_i, \quad i \in [s], \\ \boldsymbol{x} &\geq 0. \end{aligned}$$

This constraint can be added in the $E-$,$A-$,$D-$ or $\boldsymbol{c}-$optimal design problem formulations studied in this section .

Chapter 4

A Low rank reduction Theorem in Semidefinite Programming

In this chapter –which essentially recalls the work of [Sag09a]– we study the class of *semidefinite packing problems*, which encompasses as special cases some SDPs encountered in Section 3.3. The main result of this chapter is that these semidefinite packing problems admit a solution which is of low rank. A a consequence, we will see in Chapter 5 that the $\boldsymbol{c}-$ and $A-$optimal design problems reduce to a Second Order Cone Program (SOCP) which is computationally more tractable than the initial SDP; that the $E-$ optimal design problem for $K^T\boldsymbol{\theta}$ can be solved efficiently by a low-rank SDP solver when r is small (r is the number of columns of K, i.e. the number of linear functions of $\boldsymbol{\theta}$ to be estimated); and that the $D-$optimal design problem for the full parameter $\boldsymbol{\theta}$ ($K = \boldsymbol{I}$) reduces to the maximization of a geometric mean subject to SOCP constraints, which is computationally more tractable than the initial MAXDET problem.

Semidefinite packing problems were introduced by Iyengar, Phillips and Stein [IPS05]. They showed that these arise in many applications such as relaxations of combinatorial optimization problems or maximum variance unfolding, and gave an algorithm to compute approximate solutions, which is faster than the commonly used interior point methods. The problems of this class, which are the SDP analogs to the packing problems in linear programming, can be written as:

$$\begin{array}{lll} \max & \langle C, X\rangle & \\ \text{s.t.} & \langle M_i, X\rangle \leq b_i, \qquad i \in [s], & \\ & X \in \mathbb{S}^+_m, & \end{array} \tag{P_{PCK}}$$

where $C \succeq 0$, and $M_i \succeq 0,\ i \in [s]$. Our result states that when the matrix C is of rank r, Problem (P_{PCK}) has a solution that is of rank at most r (Theorem 4.1.2). In particular, when $r = 1$, the optimal SDP variable X can be factorized as $\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}}$, and we show that finding $\boldsymbol{x}$ reduces to a Second-Order Cone Program (SOCP). In this chapter, we will discuss the significance of our rank reduction theorem for the relaxations of combinatorial optimization

problems that are presented in [IPS05] (the hypothesis on the rank of the matrix C appears to be very restrictive). The consequences for the computation of optimal experimental design are the object of Chapter 5. In Section 4.2, we will extend our result to a wider class of semidefinite programs (Theorem 4.2.2), in which not all the constraints are of *packing* type. The proofs of the theorems of this chapter are given in Section 4.3.

Related work Solutions of small rank of semidefinite programs have been extensively studied over the past years. Barvinok [Bar95] and Pataki [Pat98] discovered independently that any SDP with s constraints has a solution X^* whose rank is at most

$$r^* = \left\lfloor \frac{\sqrt{8s+1}-1}{2} \right\rfloor,$$

where $\lfloor \cdot \rfloor$ denotes the integer part. This was one of the motivations of Burer and Monteiro for developing the SDPLR solver [BM03], which searches a solution of the SDP in the form $X = RR^T$, where R is a $n \times r^*$ matrix. The resulting problem is non-convex, and so the augmented Lagrangian algorithm proposed in [BM03] is not guaranteed to converge to a global optimum. However, it performs remarkably well in practice, and some conditions which ensure that the returned solution is an optimum of the SDP are provided in [BM05]. Our result shows that for a semidefinite packing problem in which the matrix C has rank r, one can force the matrix R to be of size $n \times r$ (rather than $n \times r^*$), which can lead to considerable gains in computation time when r is small.

We point out that the ratio between the optimal value of Problem (P_{PCK}) and the value of its best solution of rank one has been studied by Nemirovski, Roos, and Terlaky [NRT99]. They show that the value v^* of the SDP and the value v_1^* of its best rank-one solution satisfy:

$$v^* \geq v_1^* \geq \frac{1}{2\ln(2s\mu)} v^*, \qquad \text{where } \mu = \min(s, \max_{i \in [s]} \text{ rank } M_i). \tag{4.1}$$

This ratio can be considerably reduced in particular configurations, but to the best of our knowledge, the fact that the gap in (4.1) vanishes when the matrix C in the objective function is of rank 1 is new, except in the particular case in which every M_i is of rank 1, too [Ric08].

4.1 A rank reduction theorem

4.1.1 Main result

We start with an algebraic characterization of the semidefinite packing problems that are feasible and bounded.

Theorem 4.1.1. *Problem* (P_{PCK}) *is feasible if and only if every* b_i *is nonnegative. Moreover if Problem* (P_{PCK}) *is feasible, then this problem is bounded if and only if the range of* C *is included in the range of* $\sum_i M_i$.

The reader should note that the range inclusion condition in Theorem 4.1.1 is in fact equivalent to the feasibility of the Lagrangian dual of Problem (P_{PCK}):

$$\begin{array}{ll} \min\limits_{\boldsymbol{\mu}\geq 0} & \boldsymbol{\mu}^{\boldsymbol{T}}\boldsymbol{b} \\ \text{s.t.} & \sum\limits_i \mu_i M_i \succeq C. \end{array} \qquad (D_{\text{PCK}})$$

The main result of this chapter follows:

Theorem 4.1.2. *We assume that the conditions of Theorem 4.1.1 are fulfilled, so that Problem* (P_{PCK}) *is feasible and bounded. If* $\operatorname{rank} C = r$, *then the semidefinite packing problem* (P_{PCK}) *has a solution which is a matrix of rank at most* r.

Under a few additional conditions, we can also bound the rank of every solution. For a proof of the next statement, we refer to the last page of this chapter (proof of the second part of Theorem 4.2.2, for the case $R_i = 0$ and $\boldsymbol{h_i} = 0$ $(i \in \{0, \ldots, s\})$; note that in this case the condition $\sum_{i=1}^{s} M_i \succ 0$ is equivalent to the strict dual feasibility).

Theorem 4.1.3. *We assume that Problem* (P_{PCK}) *is feasible,* $C \neq 0$ *and* $\sum_{i=1}^{s} M_i \succ 0$. *Then, every solution* X *of Problem* (P_{PCK}) *must be of rank at most* $n - \overline{r} + r$, *where* $\overline{r} := \min\limits_{i\in[s]} \operatorname{rank}\ M_i$.

A consequence of Theorem 4.1.2 is that when the matrix in the objective function is of rank 1 $(C = \boldsymbol{c}\boldsymbol{c}^T)$, the computation of a solution X of Problem (P_{PCK}) reduces to the computation of a vector $\boldsymbol{x}$ such that $X = \boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}}$. The next result shows that this can be done very efficiently by a Second Order Cone Program (SOCP).

Corollary 4.1.4. *We assume that the conditions of Theorem 4.1.1 are fulfilled, and that* $C = \boldsymbol{c}\boldsymbol{c}^{\boldsymbol{T}}$ *for a vector* $\boldsymbol{c} \in \mathbb{R}^m$ *(i.e.* $\operatorname{rank}\ C = 1$*). Then, Problem* (P_{PCK}) *reduces to the SOCP:*

$$\begin{array}{ll} \max\limits_{\boldsymbol{x}\in\mathbb{R}^m} & \boldsymbol{c}^{\boldsymbol{T}}\boldsymbol{x} \\ s.t. & \|A_i\boldsymbol{x}\|_2 \leq \sqrt{b_i}, \qquad i = 1 \in [s], \end{array} \qquad (4.2)$$

where the matrices A_i *are such that* $M_i = A_i^T A_i$. *Moreover, if* $\boldsymbol{x}$ *is any optimal solution of Problem* (4.2), *then* $X = \boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}}$ *is an optimal solution of Problem* (P_{PCK}), *and the optimal value of* (P_{PCK}) *is* $(\boldsymbol{c}^{\boldsymbol{T}}\boldsymbol{x})^2$.

Proof. The SOCP (4.2) is simply obtained from (P_{PCK}) by substituting $\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}}$ from X and $A_i^T A_i$ from M_i. The objective function $\langle C, X\rangle$ becomes $(\boldsymbol{c}^{\boldsymbol{T}}\boldsymbol{x})^2$, and we can remove the

square by noticing that $\boldsymbol{c}^T\boldsymbol{x} \geq 0$ without loss of generality, since if $\boldsymbol{x}$ is optimal, so is $-\boldsymbol{x}$. □

In fact, the proof of Theorem 4.1.2 relies on the projection of Problem (P_{PCK}) on an appropriate subspace, which lets the reduced semidefinite packing problem be strictly feasible, as well as its dual. This reduction is not only of theoretical interest, since in some cases it may yield some important computational savings. Therefore, we next state this result as a proposition.

Let $\mathcal{I}_0 := \{i \in [s] : b_i = 0\}$ and $\mathcal{I} := [s] \setminus \mathcal{I}_0$. Let the columns of the $m \times m_0$ matrix U form an orthonormal basis of $\text{Im}(\sum_{i\in[s]} M_i)$, and the columns of the $m_0 \times m'$ matrix V form an orthonormal basis of $\text{Ker}(U^T \sum_{i\in\mathcal{I}_0} M_i U)$. We further define $C' := (UV)^T C(UV) \in \mathbb{S}^+_{m'}$ and $M'_i := (UV)^T M_i(UV) \in \mathbb{S}^+_{m'}$ (for $i \in \mathcal{I}$), and we consider the reduced problem

$$\begin{aligned} \max_{Z\in\mathbb{S}^+_{m'}} \quad & \langle C', Z\rangle & (P'_{\text{PCK}}) \\ \text{s.t.} \quad & \langle M'_i, Z\rangle \leq b_i, \qquad i \in \mathcal{I}. \end{aligned}$$

Proposition 4.1.5. *We assume that the conditions of Theorem 4.1.1 are fulfilled, so that Problem (P_{PCK}) is feasible and bounded. Then, the following properties hold:*

(i) Problem (P'_{PCK}) is strictly feasible, i.e. $\exists \overline{Z} \succ 0 :\ \forall i \in \mathcal{I}, \langle M'_i, \overline{Z}\rangle < b_i$;

(ii) The Lagrangian dual of (P'_{PCK}) is strictly feasible, i.e. $\exists \overline{\boldsymbol{\mu}} > \mathbf{0} :\ \sum_{i\in\mathcal{I}} \overline{\mu}_i M'_i \succ C'$;

(iii) If Z is a solution of Problem (P'_{PCK}), then $X := (UV)Z(UV)^T$ is an optimal solution of Problem(P_{PCK}) (which of course satisfies $\text{rank}\ X \leq \text{rank}\ Z$ and $\langle C, X\rangle = \langle C', Z\rangle$).

4.1.2 Relation with combinatorial optimization

SDP relaxations of combinatorial optimization problems have motivated the authors of [IPS05] to study semidefinite packing problems. Hence, we discuss the significance of our result for this class of problems in this section.

Semidefinite programs have been used extensively to formulate relaxations of NP-hard combinatorial optimization problems after the work of Goemans and Williamson on the approximability of MAXCUT [GW95]. These SDP relaxations often lead to optimal solutions of the related combinatorial optimization problems whenever the solution of the SDP is of small rank. As shown by Iyengar et. al. [IPS05], SDP relaxations of many combinatorial optimization problems can be cast as semidefinite packing programs. Our result therefore identifies a subclass of combinatorial optimization problems which are solvable in polynomial time. Unfortunately, this promising statement only helped us to identify trivial instances so far. For example, the MAXCUT semidefinite packing problem [IPS05] yields an exact solution of the combinatorial problem whenever it has a rank 1 solution. The matrix C in the objective function of this SDP is the Laplacian of the graph, and so it is known that

$$\text{rank}\ C = N - \kappa,$$

where N is the number of vertices and κ is the number of connected components in the graph. Our result therefore states that if a graph of N vertices has $N-1$ connected components, then it defines a MAXCUT instance that is solvable in polynomial time. Such graphs actually consist in a pair of connected vertices, plus $N-2$ isolated vertices, and the related MAXCUT instance is trivial.

Another limitation for the application of our theorem in this field is that most semidefinite packing problems arising in combinatorial optimization (including but not limited to the Lovász ϑ function SDP [Lov79] and the related Szegedy number SDP [Sze94], the vector colouring SDP [KMS98], the sparsest cut SDP [ARV09] and the sparse principal components analysis SDP [dAEJL07]) can be written in the form of (P_{PCK}), with an additional trace equality constraint $\operatorname{trace}(X) = 1$. In fact, we can show that if such an "equality constrained semidefinite packing problem" is strictly feasible, then it is equivalent to the following "classical" semidefinite packing problem:

$$\begin{aligned} \max \quad & \langle C + \lambda \boldsymbol{I}, X\rangle - \lambda \\ \text{s.t.} \quad & \langle M_i, X\rangle \leq b_i, \qquad i \in [s], \\ & \operatorname{trace} X \leq 1, \\ & X \succeq 0, \end{aligned} \tag{4.3}$$

where λ is any scalar larger than $|\lambda^*|$, where λ^* is the optimal Lagrange multiplier associated to the constraint $\operatorname{trace}(X) = 1$ (we omit the proof of this statement which is of secondary importance). Since $C + \lambda \boldsymbol{I}$ is a full rank matrix, our result does not seem to yield any valuable information for this class of problems.

4.2 Extension to "combined" problems

The proof of our main result also applies to a wider class of semidefinite programs, which can be written as:

$$\begin{aligned} \sup_{X,Y,\boldsymbol{\lambda}} \quad & \langle C, X\rangle + \langle R_0, Y\rangle + \boldsymbol{h_0}^T \boldsymbol{\lambda} \\ \text{s.t.} \quad & \langle M_i, X\rangle \leq b_i + \langle R_i, Y\rangle + \boldsymbol{h_i}^T \boldsymbol{\lambda}, \qquad i \in [s], \\ & X \in \mathbb{S}_m^+,\ Y \in \mathbb{S}_p^+,\ \boldsymbol{\lambda} \in \mathbb{R}^q, \end{aligned} \tag{P_{CMB}}$$

where **every matrix M_i and C are positive semidefinite, while the R_i are *arbitrary symmetric* matrices**. The vectors $\boldsymbol{h_i}$ are in $\mathbb{R}^q$. We denote by H the $q \times s$ matrix formed

by the columns $\boldsymbol{h_1}, \ldots, \boldsymbol{h_s}$. The Lagrangian dual of Problem (P_{CMB}) is:

$$\begin{aligned} \inf_{\boldsymbol{\mu} \geq 0} \quad & \boldsymbol{b}^T \boldsymbol{\mu} \qquad (D_{\text{CMB}}) \\ \text{s.t.} \quad & \sum_{i=1}^{s} \mu_i M_i \succeq C, \\ & R_0 + \sum_{i=1}^{s} \mu_i R_i \preceq 0. \\ & \boldsymbol{h_0} + H\boldsymbol{\mu} = \mathbf{0}. \end{aligned}$$

We have seen in Section 4.1.1 that the feasibility of both the primal (P_{PCK}) and the dual (D_{PCK}) is sufficient to guarantee that Problem (P_{PCK}) has a solution of rank at most $r := \text{rank } C$. For *combined* problems however, the feasibility of the couple of programs (P_{CMB})–(D_{CMB}) is not sufficient to guarantee the existence of a solution $(X, Y, \boldsymbol{\lambda})$ of Problem (P_{CMB}) in which $\text{rank } X \leq r$. We give indeed an example (Example 4.2.3) where the optimum in Problem (P_{CMB}) is not even attained. However, we show in the next theorem that an asymptotic result subsists. Moreover, we shall see in Theorem 4.2.2 that a solution in which X is of rank at most r exists as soon as an additional condition holds (strict dual feasibility). The proof of Theorem 4.2.2 essentially mimics that of Theorem 4.1.2 and is presented in Section 4.3.2. Theorem 4.2.1 turns out to be a consequence of Theorem 4.2.2 and is proved in Section 4.3.3.

Theorem 4.2.1. *We assume that Problems (P_{CMB}) and (D_{CMB}) are feasible. If $\text{rank } C = r$, then there exists a sequence of feasible primal variables $(X_k, Y_k, \boldsymbol{\lambda}_k)_{k \in \mathbb{N}}$ such that $\text{rank } X_k \leq r$ for all $k \in \mathbb{N}$ and $\langle C, X_k \rangle + \langle R_0, Y_k \rangle + \boldsymbol{h_0}^T \boldsymbol{\lambda}_k$ converges to the optimum of Problem (P_{CMB}) as $k \to \infty$.*

Theorem 4.2.2. *We assume that Problem (P_{CMB}) is feasible, and a refined Slater condition holds for Problem (D_{CMB}), i.e. there is a feasible dual variable which strictly satisfies the non-affine constraints:*

$$\exists \overline{\boldsymbol{\mu}} \geq \mathbf{0} : \ \sum_i \overline{\mu}_i M_i \succ C, \ R_0 + \sum_i \overline{\mu}_i R_i \prec 0, \ \boldsymbol{h_0} + H\overline{\boldsymbol{\mu}} = \mathbf{0}.$$

If $\text{rank } C = r$, then Problem (P_{CMB}) has a solution $(X, Y, \boldsymbol{\lambda})$ in which $\text{rank } X \leq r$. Moreover, if $C \neq 0$, then every solution $(X, Y, \boldsymbol{\lambda})$ of Problem (P_{CMB}) is such that $\text{rank } X \leq n - \overline{r} + r$, where $\overline{r} := \min_{i \in [s]} \text{rank } M_i$.

As in the previous section, we have a result of reduction to a SOCP, which holds when C is of rank 1, every $R_i = 0$ and $\boldsymbol{h_0} = \mathbf{0}$. Recall that H denotes the matrix formed by the columns $\boldsymbol{h_1}, \ldots, \boldsymbol{h_s}$.

Corollary 4.2.4. *Consider the following "combined" semidefinite packing problem:*

$$\begin{array}{rll} \sup\limits_{X \in \mathbb{S}_m,\ \boldsymbol{\lambda} \in \mathbb{R}^q} & \langle C, X \rangle & \\ s.t. & \langle M_i, X \rangle \leq \boldsymbol{h_i}^T \boldsymbol{\lambda} + b_i, & i \in [s], \\ & X \succeq 0. & \end{array} \tag{4.5}$$

Assume that $C = \boldsymbol{c}\boldsymbol{c}^{\boldsymbol{T}}$ has rank 1. If Problem (4.5) and its Lagrangian dual are feasible, i.e.

(i) $\exists \overline{\boldsymbol{\lambda}} \in \mathbb{R}^q : \ H^T \overline{\boldsymbol{\lambda}} + \boldsymbol{b} \geq 0$;

(ii) $\exists \overline{\boldsymbol{\mu}} \geq \mathbf{0} : \ \sum_i \overline{\mu}_i M_i \succeq C, \ \boldsymbol{h_0} + H\overline{\boldsymbol{\mu}} = \mathbf{0}$,

then, Problem (4.5) is bounded, and its optimal value is the square of the optimal value of the following SOCP:

$$\begin{array}{rll} \sup\limits_{\boldsymbol{x} \in \mathbb{R}^m,\ \boldsymbol{\lambda} \in \mathbb{R}^q} & \boldsymbol{c}^T \boldsymbol{x} & \\ s.t. & \left\| \begin{bmatrix} 2A_i \boldsymbol{x} \\ \boldsymbol{h_i}^T \boldsymbol{\lambda} + b_i - 1 \end{bmatrix} \right\|_2 \leq \boldsymbol{h_i}^T \boldsymbol{\lambda} + b_i + 1, & i \in [s], \end{array} \tag{4.6}$$

where the matrices A_i are such that $M_i = A_i^T A_i$. Moreover, if $(\boldsymbol{x}, \boldsymbol{\lambda})$ is a solution of Problem (4.6), then $(\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}}, \boldsymbol{\lambda})$ is a solution of Problem (4.5), and the optimal value of (4.5) is $(\boldsymbol{c}^T \boldsymbol{x})^2$.

Proof. Theorem 4.2.1 guarantees the existence of a sequence of feasible variables $(X_k, \boldsymbol{\lambda}_k)_{k \in \mathbb{N}}$ in which X_k has rank 1, i.e. $X_k = \boldsymbol{x_k}\boldsymbol{x_k}^T$, and $\langle C, X_k \rangle = (\boldsymbol{c}^T \boldsymbol{x}_k)^2$ converges to the optimum of Problem (4.5). This optimal value is therefore equal to the supremum

---

Example 4.2.3. Consider the following *combined* semidefinite packing problem:

$$\begin{array}{rll} \sup\limits_{X \in \mathbb{S}_2^+,\ \boldsymbol{\lambda} \in \mathbb{R}^2} & \frac{3}{100} \left\langle \begin{pmatrix} 81 & 9 \\ 9 & 1 \end{pmatrix}, X \right\rangle - \lambda_1 - 3\lambda_2 & \\ \text{s.t.} & 0 \leq 1 + \lambda_1 & \\ & X_{1,1} \leq 1 + \lambda_2 & \\ & X_{2,2} \leq 1 + 3\lambda_1 + \lambda_2. & \end{array} \tag{4.4}$$

This problem is in the form of (P_{CMB}) indeed, with $C = \boldsymbol{c}\boldsymbol{c}^T$, $\boldsymbol{c} = \frac{\sqrt{3}}{10}[9 \quad 1]^T$, $\boldsymbol{h_0} = [-1 \quad -3]^T$,

$$M_1 = 0, \ M_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \ M_3 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad H = \begin{pmatrix} 1 & 0 & 3 \\ 0 & 1 & 1 \end{pmatrix}.$$

Problem (4.4) is clearly feasible (e.g. for $X = 0$, $\boldsymbol{\lambda} = \mathbf{0}$), and the reader can verify that $\boldsymbol{\mu} = \frac{1}{10}[1 \quad 27 \quad 3]^T$ is dual feasible (in fact, this is the only dual feasible vector, and hence the dual problem does not satisfy the Slater constraints qualification). The value of the optimum is $\frac{31}{10}$, and can be approached arbitrarily closely for the sequence of feasible variables $(\boldsymbol{x}_k \boldsymbol{x}_k^T, \boldsymbol{\lambda_k})_{k \in \mathbb{N}}$, where for all $k \geq 0$, $\boldsymbol{x}_k = [\sqrt{3+k} \quad \sqrt{k}]^T$, $\boldsymbol{\lambda}_k = [-1 \quad k+2]^T$, while this optimum is not attained by any couple $(X, \boldsymbol{\lambda})$ of (bounded) feasible variables.

of $(\boldsymbol{c}^T\boldsymbol{x})^2$, over all the pairs of vectors $(\boldsymbol{x}, \boldsymbol{\lambda}) \in \mathbb{R}^m \times \mathbb{R}^q$ such that $(\boldsymbol{x}\boldsymbol{x}^T, \boldsymbol{\lambda})$ is feasible for Problem (4.5). As in the proof of Corollary 4.1.4, we notice that if $(\boldsymbol{x}\boldsymbol{x}^T, \boldsymbol{\lambda})$ is feasible for Problem (4.5), so is $((-\boldsymbol{x})(-\boldsymbol{x})^T, \boldsymbol{\lambda})$, hence we can remove the square in the objective function.

The SOCP (4.6) is simply obtained from (4.5) by substituting $\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}}$ from X and $A_i^T A_i$ from M_i. We also used the fact that for any vector $\boldsymbol{z}$ and for any scalar α, the hyperbolic constraint

$$\|\boldsymbol{z}\|_2^2 \leq \alpha$$

is equivalent to the second order cone constraint

$$\left\| \begin{bmatrix} 2\boldsymbol{z} \\ \alpha - 1 \end{bmatrix} \right\|_2 \leq \alpha + 1.$$

□

4.3 Proofs of the theorems

4.3.1 Results of Section 4.1.1

Proof of Theorem 4.1.1. The fact that Problem (P_{PCK}) is feasible if and only if every b_i is nonnegative is clear, since $X = 0$ is always feasible in this case and $M_i \succeq 0, X \succeq 0$, implies $\langle M_i, X \rangle \geq 0$.

Now, we assume that each b_i is nonnegative, and we show that Problem (P_{PCK}) is bounded if and only if $\operatorname{Im} C \subset \operatorname{Im} \sum_i M_i$. The positive semidefiniteness of the matrices M_i implies that there exists matrices A_i $(i \in [s])$ such that $A_i^T A_i = M_i$, and $[A_1^T, \cdots, A_s^T][A_1^T, \cdots, A_s^T]^T = \sum_i M_i$. We also consider a decomposition $C = \sum_{k=1}^r \boldsymbol{c_k}\boldsymbol{c_k}^T$. For any factorization $M = A^T A$ of a positive semidefinite matrix M, it is known that $\operatorname{Im} M = \operatorname{Im} A$, and so the following equivalence relations hold:

$$\begin{aligned} \operatorname{Im} C \subset \operatorname{Im} \sum_i M_i &\iff \forall k \in [r],\ \boldsymbol{c_k} \in \operatorname{Im}(\sum_i M_i) = \operatorname{Im}([A_1^T, \cdots, A_s^T]) \\ &\iff \forall k \in [r],\ \boldsymbol{c_k} \in \left(\bigcap_{i=1}^s \operatorname{Ker}(A_i) \right)^{\perp}. \end{aligned} \tag{4.7}$$

We first assume that the range inclusion condition does not hold. Relation (4.7) shows that

$$\exists k \in [r], \exists \boldsymbol{h} \in \mathbb{R}^m : \forall i \in [s], \quad A_i \boldsymbol{h} = 0, \quad \boldsymbol{c_k}^T \boldsymbol{h} \neq 0.$$

Now, notice that $X = \alpha \boldsymbol{h}\boldsymbol{h}^T$ is feasible for all $\alpha > 0$, since $\alpha \langle A_i^T A_i, \boldsymbol{h}\boldsymbol{h}^T \rangle = 0 \leq b_i$. This contradicts the fact that Problem (P_{PCK}) is bounded, because $\langle C, X \rangle \geq \alpha (\boldsymbol{c_k}^T \boldsymbol{h})^2$, and α can be chosen arbitrarily large.

Conversely, if the range inclusion holds, we consider the Lagrangian dual (D_{PCK}) of Problem (P_{PCK}): The range inclusion condition indicates that this problem is feasible, because it implies the existence of a scalar $\lambda > 0$ such that $\lambda \sum_i M_i \succeq C$ (we point out that a convenient value for λ is $\sum_{k=1}^r \boldsymbol{c_k}^T (\sum_i M_i)^\dagger \boldsymbol{c_k}$; this can be seen with the help of the Schur complement lemma). This means that Problem (D_{PCK}) has a finite optimal value $OPT \leq \lambda \sum_i b_i$, and by weak duality, Problem (P_{PCK}) is bounded (its optimal value cannot exceed OPT). □

Before proving Theorem 4.1.2, we need to show that we can project Problem (P_{PCK}) on a subspace such that the projected problem (P'_{PCK}) and its Lagrangian dual are strictly feasible (Proposition 4.1.5).

Proof of Proposition 4.1.5. Let $\mathcal{I}_0, \mathcal{I}, U$ and V be defined as in the paragraph preceding the statement of the proposition (page 68). Note that every matrix M_i can be decomposed as $M_i = U\tilde{M}_i U^T$ for a given matrix $\tilde{M}_i$, because its range is included in the range of $\sum_i M_i$ (we have $\tilde{M}_i = U^T M_i U$). The same observation holds for C, which can be decomposed as $C = U\tilde{C}U^T$ (we have assumed the range inclusion $\operatorname{Im} C \subset \operatorname{Im} \sum_i M_i$). Hence, Problem ($P_{\text{PCK}}$) is equivalent to:

$$\begin{aligned} \max_{X \succeq 0} \quad & \langle \tilde{C}, U^T X U \rangle \\ \text{s.t.} \quad & \langle \tilde{M}_i, U^T X U \rangle \leq b_i, \qquad i \in [s]. \end{aligned}$$

After the change of variable $Z_0 = U^T X U$ (Z_0 is a positive semidefinite matrix if X is), we obtain a reduced semidefinite packing problem

$$\begin{aligned} \max_{Z_0 \succeq 0} \quad & \langle \tilde{C}, Z_0 \rangle \\ \text{s.t.} \quad & \langle \tilde{M}_i, Z_0 \rangle \leq b_i, \qquad i \in [s]. \end{aligned} \tag{4.8}$$

By construction, if Z_0 is a solution of (4.8), then $X := U Z_0 U^T$ is a solution of (P_{PCK}). Note that the projected matrices in the constraints now satisfy $\sum_i \tilde{M}_i = U^T (\sum_i M_i) U \succ 0$.

We shall now consider a second projection, in order to get rid of the constraints in which $b_i = 0$. Note that each constraint indexed by $i \in \mathcal{I}_0$ is equivalent to imposing that Z_0 belongs to the nullspace of the matrix $\tilde{M}_i$. Since the columns of V form a basis of $\cap_{i \in \mathcal{I}_0} \operatorname{Ker} \tilde{M}_i$, any semidefinite matrix Z_0 which is feasible for Problem (4.8) must be of the form VZV^T for some positive semidefinite matrix Z. Hence, Problem (4.8) reduces to:

$$\begin{aligned} \max_{Z \succeq 0} \quad & \langle V^T \tilde{C} V, Z \rangle \\ \text{s.t.} \quad & \langle V^T \tilde{M}_i V, Z \rangle \leq b_i, \qquad i \in \mathcal{I}. \end{aligned} \tag{4.9}$$

which is nothing but Problem (P'_{PCK}), because $V^T \tilde{M}_i V = V^T U^T M_i U V = M'_i$ and $V^T \tilde{C} V = C'$. By construction, If Z is a solution of (4.9)≡(P'_{PCK}), then VZV^T is a

solution of (4.8), and $(UV)Z(UV)^T$ is a solution of the original problem (P_{PCK}). This proves the point (iii) of the proposition.

We have pointed out above that $\sum_i \tilde{M}_i \succ 0$. Therefore, there exists a real $\lambda > 0$ such that $\lambda \sum_i \tilde{M}_i \succ \tilde{C}$, and $\lambda \sum_i M'_i = V^T\big(\lambda \sum_i \tilde{M}_i\big)V \succ V^T\tilde{C}V = C'$. This proves the strict dual feasibility of Problem (P'_{PCK}) (point (ii) of the proposition). Finally, since every b_i is positive for $i \in \mathcal{I}$, it is clear that the matrix $\overline{Z} = \varepsilon \boldsymbol{I} \succ 0$ is strictly feasible for Problem (P'_{PCK}) as soon as $\varepsilon > 0$ is sufficiently small. This establishes the point (i), and the proposition is proved. □

We can now prove the main result of this chapter. In fact, Theorem 4.1.2 can be derived from the extension to combined problems (Theorem 4.2.2), but this would somehow hide the fact that the proof is much simpler in the "non-combined case". Therefore we provide the proofs of these two similar results separately.

We will first show that the result holds when every M_i is positive definite, thanks to the complementary slackness relation. Then, the general result is obtained by continuity. We point out at the end of this section the sketch of an alternative proof of Theorem 4.1.2 for the case in which $r = 1$, based on the bidual of Problem (P_{PCK}) and Schur complements, that shows directly that Problem (P_{PCK}) reduces to the SOCP (4.2).

Proof of Theorem 4.1.2. We will show that the result of the theorem holds for any semidefinite packing problem which is strictly feasible, and whose dual is strictly feasible. Then, by Proposition 4.1.5, we can say that Problem (P'_{PCK}) has a solution Z of rank at most $r' := \text{rank } C'$, and $X := (UV)^T Z (UV)$ is a solution of the original problem which is of rank at most $r' \leq r$.

So let us assume without loss of generality that (P_{PCK}) and (D_{PCK}) are strictly feasible:

$$\forall i \in [s], b_i > 0 \quad \text{and} \quad \exists \lambda > 0 : \lambda \sum_i M_i \succ C.$$

The Slater condition is fulfilled for this pair of programs, and so strong duality holds (the optimal value of (P_{PCK}) equals the optimal value of (D_{PCK})), and the dual problem attains its optimum. In addition, the strict dual feasibility implies that (P_{PCK}) also attains its optimum. The pairs of primal and dual solutions $(X^*, \boldsymbol{\mu}^*)$ are characterized by the Karush-Kuhn-Tucker (KKT) conditions:

$$\begin{aligned}
\text{Primal Feasibility:} \quad & \forall i \in [s], \quad \langle M_i, X^* \rangle \leq b_i; \\
& X^* \succeq 0; \\
\text{Dual Feasibility:} \quad & \boldsymbol{\mu}^* \geq 0, \quad \sum_{i=1}^{s} \mu_i^* M_i \succeq C; \\
\text{Complementary Slackness:} \quad & \Big(\sum_{i=1}^{s} \mu_i^* M_i - C\Big)\, X^* = 0, \\
& \forall i \in [s], \ \mu_i^*(b_i - \langle M_i, X^* \rangle) = 0.
\end{aligned}$$

Now, we consider the case in which $M_i \succ 0$ for all i, and we choose an arbitrary pair of primal and dual optimal solutions $(X^*, \boldsymbol{\mu}^*)$. The dual feasibility relation implies $\boldsymbol{\mu}^* \neq \mathbf{0}$, and so $\sum_i \mu_i^* M_i$ is a positive definite matrix (we exclude the trivial case $C = 0$). Since C is of rank r, we deduce that

$$\operatorname{rank}(\sum_i \mu_i^* M_i - C) \geq n - r.$$

Finally, the complementary slackness relation indicates that the columns of X^* belong to the nullspace of $(\sum_i \mu_i^* M_i - C)$, which is a vector space of dimension at most $n - (n - r) = r$, and so we conclude that $\operatorname{rank} X^* \leq r$.

We now turn to the study of the general case in which $M_i \succeq 0$. To this end, we consider the perturbed problems

$$\begin{array}{rl} \max & \langle C, X \rangle \\ \text{s.t.} & \langle M_i + \varepsilon \boldsymbol{I}, X \rangle \leq b_i \\ & X \succeq 0, \end{array} \qquad (P_\varepsilon)$$

and

$$\begin{array}{rl} \min\limits_{\boldsymbol{\mu} \geq 0} & \sum\limits_{i=1}^{s} \mu_i b_i, \\ \text{s.t.} & \sum\limits_{i=1}^{s} \mu_i (M_i + \varepsilon \boldsymbol{I}) \succeq C. \end{array} \qquad (D_\varepsilon)$$

where $\varepsilon \geq 0$. Note that the strict feasibility of the unperturbed problems (P_{PCK}) and (D_{PCK}) implies that of (P_ε) and (D_ε) on a neighborhood $\varepsilon \in [0, \varepsilon_0]$, $\varepsilon_0 > 0$. We denote by $(X^\varepsilon, \boldsymbol{\mu}^\varepsilon)$ a pair of primal and dual solutions of (P_ε)–(D_ε).

If $\varepsilon > 0$, $M_i + \varepsilon \boldsymbol{I} \succ 0$ and it follows from the previous discussion that X^ε is of rank at most r. We show below that we can choose the optimal variables $(X^\varepsilon, \boldsymbol{\mu}^\varepsilon)_{\varepsilon \in]0, \varepsilon_0]}$ within a bounded region, so that we can construct a converging subsequence $(X^{\varepsilon_k}, \boldsymbol{\mu}^{\boldsymbol{\varepsilon_k}})_{k \in \mathbb{N}},\ \varepsilon_k \to 0$ from these variables. To conclude, we will see that the limit $(X^0, \boldsymbol{\mu}^{\mathbf{0}})$ satisfies the KKT conditions for Problems (P_{PCK})–(D_{PCK}), and that X^0 is of rank at most r.

Let us denote the optimal value of Problems (P_ε)–(D_ε) by $OPT(\varepsilon)$. Since the constraints of the primal problem becomes tighter when ε grows, it is clear that $OPT(\varepsilon)$ is nonincreasing with respect to ε, so that

$$\forall \varepsilon \in [0, \varepsilon_0],\ OPT(\varepsilon_0) \leq OPT(\varepsilon) \leq OPT(0).$$

We have:

$$\lambda(\sum_i M_i + \varepsilon \boldsymbol{I}) - C \succ \lambda(\sum_i M_i) - C,$$

and so we can write

$$\begin{aligned}\langle \lambda \sum_i M_i - C, X^\varepsilon \rangle &\leq \langle \lambda \sum_i (M_i + \varepsilon \boldsymbol{I}) - C, X^\varepsilon \rangle \\ &= \lambda \langle \sum_i (M_i + \varepsilon \boldsymbol{I}), X^\varepsilon \rangle - OPT(\varepsilon) \\ &\leq \lambda \sum_i b_i - OPT(\varepsilon_0)\end{aligned}$$

where the equality comes from the expression of $OPT(\varepsilon)$ and the latter inequality follows from the constraints of the Problem (P_ε). The matrix $\lambda \sum_i M_i - C$ is positive definite by assumption and its smallest eigenvalue λ' is therefore positive. Hence,

$$\lambda' \text{ trace } X^\varepsilon \leq \langle \lambda \sum_i M_i - C, X^\varepsilon \rangle \leq \overline{\boldsymbol{\mu}}^T \boldsymbol{b} - OPT(\varepsilon) \leq \lambda \sum_i b_i - OPT(\varepsilon_0).$$

This shows that the positive semidefinite matrix X^ε has its trace bounded, and therefore all its entries are bounded.

It remains to show that the dual optimal variable $\boldsymbol{\mu}^\varepsilon \geq \mathbf{0}$ is bounded. This is simply done by writing:

$$\forall i \in [s], \quad b_i \mu_i^\varepsilon \leq \boldsymbol{b}^T \boldsymbol{\mu}^\varepsilon = OPT(\varepsilon) \leq OPT(0).$$

By assumption, $b_i > 0$, and the entries of the vector $\boldsymbol{\mu}^\varepsilon \geq \mathbf{0}$ are bounded.

We can therefore construct a sequence of pairs of primal and dual optimal solutions $(X^\varepsilon, \boldsymbol{\mu}^{\boldsymbol{\varepsilon_k}})_{k \in \mathbb{N}}$ that converges, with $\varepsilon_k \underset{k \to \infty}{\longrightarrow} 0$, $\varepsilon_k > 0$. The limit X^0 of this sequence is of rank at most r, because the rank is a lower semicontinuous function and $\text{rank } X^{\varepsilon_k} \leq r$ for all $k \in \mathbb{N}$. It remains to show that X^0 is a solution of Problem (P_{PCK}). The $\varepsilon-$perturbed KKT conditions must hold for all $k \in \mathbb{N}$, and so they hold for the pair $(X_0, \boldsymbol{\mu^0})$ by taking the limit (the limit of any sequence of positive semidefinite matrices is a positive semidefinite matrix because $\mathbb{S}_m^+$ is closed). This concludes the proof. □

Sketch of an alternative proof of Theorem 4.1.2 when $r = 1$

Proof. By Proposition 4.1.5, we only need to show that the result holds for the reduced problem (P'_{PCK}), and so we assume without loss of generality that strong duality holds for all the optimization problems considered below.

When $r = 1$, there is a vector $\boldsymbol{c}$ such that $C = \boldsymbol{cc^T}$ and the dual problem of (P_{PCK}) takes the form:

$$\begin{aligned}\min_{\boldsymbol{\mu} \geq 0} \quad & \boldsymbol{\mu^T b} \\ \text{s.t.} \quad & \boldsymbol{cc^T} \preceq \sum_i \mu_i M_i.\end{aligned} \tag{4.10}$$

Now, setting $t = \boldsymbol{\mu}^{\boldsymbol{T}}\boldsymbol{b}$, and $\boldsymbol{w} = \frac{\boldsymbol{\mu}}{t}$, so that the new variable $\boldsymbol{w}$ satisfies $\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{b} = 1$, the constraint of the previous problem becomes $\frac{\boldsymbol{c}\boldsymbol{c}^{\boldsymbol{t}}}{t} \preceq \sum_i w_i M_i$. This matrix inequality, together with the fact that the optimal t is positive, can be reformulated thanks to the Schur complement lemma, and (4.10) is equivalent to:

$$\begin{aligned} \min_{t\in\mathbb{R},\boldsymbol{w}\geq\mathbf{0}} \quad & t \\ \text{s.t.} \quad & \left(\begin{array}{c|c} \sum_i w_i M_i & \boldsymbol{c} \\ \hline \boldsymbol{c}^T & t \end{array}\right) \succeq 0. \\ & \boldsymbol{w}^T\boldsymbol{b} = 1. \end{aligned} \tag{4.11}$$

We dualize this SDP once again to obtain the bidual of Program (P_{PCK}) (strong duality holds):

$$\begin{aligned} \max_{\beta\in\mathbb{R},Z\in\mathbb{S}^+_{m+1}} \quad & -\beta - 2\boldsymbol{v}^T\boldsymbol{c} \\ \text{s.t.} \quad & \langle W, M_i\rangle \leq \beta b_i, \quad i \in [s] \\ & Z = \left(\begin{array}{c|c} W & \boldsymbol{v} \\ \hline \boldsymbol{v^T} & 1 \end{array}\right) \succeq 0. \end{aligned} \tag{4.12}$$

We notice that the last matrix inequality is equivalent to $W \succeq \boldsymbol{v}\boldsymbol{v^T}$, using a Schur complement. Since $M_i \succeq 0$, we can assume that $W = \boldsymbol{v}\boldsymbol{v^T}$ without loss of generality, and (4.12) becomes:

$$\begin{aligned} \max_{\beta\in\mathbb{R},\boldsymbol{v}\in\mathbb{R}^m} \quad & -\beta - 2\boldsymbol{v}^T\boldsymbol{c} \\ \text{s.t.} \quad & \|A_i\boldsymbol{v}\|^2 \leq \beta b_i, \quad i \in [s], \end{aligned} \tag{4.13}$$

where A_i is a matrix such that $A_i^T A_i = M_i$.

We now define the new variables $\alpha = \sqrt{\beta}$, and $\boldsymbol{x} = \frac{\boldsymbol{v}}{\alpha}$, so that (4.13) becomes:

$$\begin{aligned} \max_{\boldsymbol{x}\in\mathbb{R}^m} \quad & \left(\max_{\alpha} -\alpha^2 - 2\alpha\boldsymbol{x^T}\boldsymbol{c}\right) \\ \text{s.t.} \quad & \|A_i\boldsymbol{x}\| \leq \sqrt{b_i}, \quad i = 1 \in [s]. \end{aligned} \tag{4.14}$$

The reader can finally verify that the value of the max within parenthesis is $(\boldsymbol{c^T}\boldsymbol{x})^2$, and we have proved that the SDP (P_{PCK}) reduces to the SOCP (4.2). By the way, this guarantees that the SDP (P_{PCK}) has a rank-one solution. □

4.3.2 Proof of Theorem 4.2.2

Before we give the proof of Theorem 4.2.2, we need one additional technical lemma, which shows that one can assume without loss of generality that the primal problem is strictly feasible, and that the vector space spanned by the vectors $\boldsymbol{h_0}, \boldsymbol{h_1}, \ldots, \boldsymbol{h_s}$ coincides

with the cone generated by the same vectors. One can consider this lemma as the analog of Proposition 4.1.5 for combined problems.

Lemma 4.3.1. *We assume that the conditions of Theorem 4.2.2 are fulfilled. Then, there exists a subset $\mathcal{I} \subset [s]$, as well as matrices $C' \succeq 0$ and $M'_i \succeq 0$ $(i \in \mathcal{I})$, so that the reduced "combined" semidefinite packing problem*

$$\max_{Z \succeq 0,\ Y \succeq 0,\ \boldsymbol{\lambda}} \langle C', Z\rangle + \langle R_0, Y\rangle + \boldsymbol{h_0}^T\boldsymbol{\lambda} \qquad s.t. \qquad \forall i \in \mathcal{I},\ \langle M'_i, Z\rangle \leq b_i + \langle R_i, Y\rangle + \boldsymbol{h_i}^T\boldsymbol{\lambda}$$

has the same optimal value as (P_{CMB}) and satisfies the following properties:

- (i) $\exists (Z' \succ 0, Y' \succ 0, \boldsymbol{\lambda'}) :\ \forall i \in \mathcal{I},\ \langle M_i, Z'\rangle < b_i + \langle R_i, Y'\rangle + \boldsymbol{h_i}^T\boldsymbol{\lambda'}$;
- (ii) *The cone K generated by the vectors $(\boldsymbol{h_i})_{i\in\{0\}\cup\mathcal{I}}$ is a vector space.*
- (iii) $\mathrm{rank}\ C' \leq \mathrm{rank}\ C$;
- (iv) *There is a matrix U with orthonormal columns such that if $(Z, Y, \boldsymbol{\lambda})$ is a solution of the reduced problem, then $(X := UZU^T, Y, \boldsymbol{\lambda})$ is a solution of Problem (P_{CMB}) (which of course satisfies $\mathrm{rank}\ X \leq \mathrm{rank}\ Z$).*

Proof. In this lemma, (i) and (ii) are the properties that we will need to prove Theorem 4.2.2. Properties (iii) and (iv) ensure that if the theorem holds for the reduced problem, then the result also holds for the initial problem (P_{CMB}). We handle separately the cases in which the initial problem does not satisfy the property (i) or (ii). If both cases arise simultaneously, we obtain the result of this lemma by applying successively the following two reductions.

Let $(X^*, Y^*, \boldsymbol{\lambda}^*)$ be an optimal solution of Problem (P_{CMB}) ; the existence of a solution is guaranteed by the (refined) Slater condition satisfied by the dual problem indeed (see e.g. [Roc70, Ber95]). We denote by $\mathcal{I}_0 \subset [s]$ the subset of indices for which $b_i + \langle R_i, Y^*\rangle + \boldsymbol{h_i}^T\boldsymbol{\lambda^*} = 0$ (note that we have $b_i + \langle R_i, Y^*\rangle + \boldsymbol{h_i}^T\boldsymbol{\lambda^*} \geq 0$ for all i because $M_i \succeq 0$ implies $\langle M_i, X^*\rangle \geq 0$). We define $\mathcal{I} := [s] \setminus \mathcal{I}_0$. In Problem (P_{CMB}), we can replace the constraint $\langle M_i, X\rangle \leq b_i + \langle R_i, Y\rangle + \boldsymbol{h_i}^T\boldsymbol{\lambda}$ by $\langle M_i, X\rangle = 0$ for all $i \in \mathcal{I}_0$, since $(X^*, Y^*, \boldsymbol{\lambda^*})$ satisfies this stronger set of constraints. For a feasible positive semidefinite matrix X, this implies $\langle \sum_{i\in\mathcal{I}_0} M_i, X\rangle = 0$, and even $\sum_{i\in\mathcal{I}_0} M_i X = 0$. Therefore, X is of the form UZU^T for some positive semidefinite matrix Z, where the columns of U form an orthonormal basis of the nullspace of $M_0 := \sum_{i\in\mathcal{I}_0} M_i$ (U is obtained by taking the eigenvectors corresponding to the vanishing eigenvalues of M_0). Hence, Problem (P_{CMB}) is equivalent to:

$$\begin{aligned} \max \quad & \langle U^TCU, Z\rangle + \langle R_0, Y\rangle + \boldsymbol{h_0}^T\boldsymbol{\lambda} \\ \text{s.t.} \quad & \langle U^TM_iU, Z\rangle \leq b_i + \langle R_i, Y\rangle + \boldsymbol{h_i}^T\boldsymbol{\lambda}, \qquad i \in \mathcal{I}, \\ & Z \succeq 0,\ Y \succeq 0. \end{aligned} \tag{4.15}$$

We have thus reduced the problem to one for which $b_i + \langle R_i, Y^*\rangle + \boldsymbol{h_i}^T\boldsymbol{\lambda^*} > 0$ for all i, and strict feasibility follows (i.e. property (i) holds, consider $\boldsymbol{\lambda'} = \boldsymbol{\lambda^*}, Y' = Y^* + \eta_1\boldsymbol{I}$, and $Z' = \eta_2\boldsymbol{I}$ for sufficiently small reals $\eta_1 > 0$ and $\eta_2 > 0$). Moreover, the projected

matrix $C' := U^T C U$ in the objective function has a smaller rank than C (i.e. (iii) holds). Finally, (iv) holds for the reduced problem by construction: if $(Z, Y, \boldsymbol{\lambda})$ is a solution of Problem (4.15), then $(X := UZU^T, Y, \boldsymbol{\lambda})$ is a solution of Problem (P_{CMB}), both problems have the same optimal value, and of course $\text{rank } X \leq \text{rank } Z$.

We now handle the second case, in which Property (ii) does not hold for Problem (P_{CMB}). The set $K = \{\ [\boldsymbol{h_0}, H]\boldsymbol{v},\ \boldsymbol{v} \in \mathbb{R}^{s+1}, \boldsymbol{v} \geq \boldsymbol{0}\}$ is a closed convex cone. Hence, it is known that it can be decomposed as $K = L + Q$, where L is a vector space and $Q \subset L^\perp$ is a closed convex pointed cone ($L = K \cap (-K)$ is the *lineality space* of K). The interior of the dual cone Q^* is therefore nonempty, i.e. $\exists \boldsymbol{\lambda} : \forall \boldsymbol{q} \in Q \setminus \{\boldsymbol{0}\}, \boldsymbol{\lambda}^T \boldsymbol{q} > 0$. Let $\boldsymbol{\lambda_0}$ be the orthogonal projection of $\boldsymbol{\lambda}$ on $L^\perp$, so that $\boldsymbol{\lambda_0}^T \boldsymbol{q} = \boldsymbol{\lambda}^T \boldsymbol{q} > 0$ for all $\boldsymbol{q} \in Q \setminus \{\boldsymbol{0}\}$, and $\boldsymbol{\lambda_0}^T \boldsymbol{x} = 0$ for all $\boldsymbol{x} \in L$. Now, we define the set of indices $\mathcal{I} = \{i \in [s] : \boldsymbol{h_i} \in L\}$, and its complement $\mathcal{I}_0 = [s] \setminus \mathcal{I}$. For all $i \in \mathcal{I}_0$, $\boldsymbol{h_i} = \boldsymbol{x_i} + \boldsymbol{q_i}$ for a vector $\boldsymbol{x_i} \in L$ and a vector $\boldsymbol{q_i} \in Q \setminus \{\boldsymbol{0}\}$, so that $\boldsymbol{\lambda_0}^T \boldsymbol{h_i} = \boldsymbol{\lambda_0}^T \boldsymbol{x_i} + \boldsymbol{\lambda_0}^T \boldsymbol{q_i} = \boldsymbol{\lambda_0}^T \boldsymbol{q_i} > 0$. For the indices $i \in \mathcal{I}$, it is clear that $\boldsymbol{\lambda_0}^T \boldsymbol{h_i} = 0$. Finally, since $\boldsymbol{h_0} + H\overline{\boldsymbol{\mu}} = 0$, we have $-\boldsymbol{h_0} \in K$, so that $\boldsymbol{h_0} \in L$ and $\boldsymbol{h_0}^T \boldsymbol{\lambda} = 0$. To sum up, we have proved the existence of a vector $\boldsymbol{\lambda_0}$ for which

$$\forall i \in \{0\} \cup \mathcal{I},\ \boldsymbol{\lambda_0}^T \boldsymbol{h_i} = 0 \quad \text{and} \quad \forall i \in \mathcal{I}_0, \boldsymbol{\lambda_0}^T \boldsymbol{h_i} > 0.$$

Let $(X^*, Y^*, \boldsymbol{\lambda}^*)$ be an optimal solution of Problem (P_{CMB}). For all positive real t, $(X^*, Y^*, \boldsymbol{\lambda}^* + t\boldsymbol{\lambda_0})$ is also a solution, because it is feasible and has the same objective value. Letting $t \to \infty$, we see that the constraints of the problem that are indexed by $i \in \mathcal{I}_0$ may be removed without changing the optimum. We have thus reduced the problem to one for which (ii) holds.

□

We can now prove Theorem 4.2.2. The proof mimics that of Theorem 4.1.2, i.e. we first show that the result holds when each M_i is positive definite, and the general result is obtained by continuity. The only difference is how we show that we can choose optimal variables $(X^\varepsilon, Y^\varepsilon, \boldsymbol{\lambda}^\varepsilon, \boldsymbol{\mu}^\varepsilon)_{\varepsilon \in]0, \varepsilon_0]}$ for a perturbed problem within a bounded region.

Proof of Theorem 4.2.2. By Lemma 4.3.1, we may assume without loss of generality that $K = \text{cone}\{\boldsymbol{h_0}, \ldots, \boldsymbol{h_s}\} \supset -K$ and that the primal problem is strictly feasible. The strict feasibility of the primal problem ensures that strong duality holds, i.e. the optimal value of (P_{CMB}) equals the optimal value of (D_{CMB}), and the optimum is attained in the dual problem. Moreover, the (refined) Slater constraints qualification for the dual problem guarantees the existence of primal optimal variables as well (see e.g. Theorem 28.2 in [Roc70]). The pairs of primal and dual solutions $\left((X^*, Y^*, \boldsymbol{\lambda}^*), \boldsymbol{\mu}^*\right)$ are characterized by the Karush-

Kuhn-Tucker (KKT) conditions:

$$\begin{aligned}
\text{Primal Feasibility:} \quad & \forall i \in [s], \quad \langle M_i, X^* \rangle \leq b_i + \langle R_i, Y^* \rangle + \boldsymbol{h_i}^T \boldsymbol{\lambda}^*, \\
& X^* \succeq 0, \ Y^* \succeq 0; \\
\text{Dual Feasibility:} \quad & \boldsymbol{\mu}^* \geq 0, \quad \sum_{i=1}^{s} \mu_i^* M_i \succeq C, \\
& R_0 + \sum_{i=1}^{s} \mu_i^* R_i \preceq 0, \quad \boldsymbol{h_0} + H\boldsymbol{\mu}^* = 0; \\
\text{Complementary Slackness:} \quad & (\sum_{i=1}^{s} \mu_i^* M_i - C)\ X^* = 0, \qquad (R_0 + \sum_{i=1}^{s} \mu_i^* R_i)\ Y^* = 0, \\
& \forall i \in [s], \ \mu_i^* (b_i + \langle R_i, Y^* \rangle + \boldsymbol{h_i}^T \boldsymbol{\lambda}^* - \langle M_i, X^* \rangle) = 0.
\end{aligned}$$

Now, we consider the case in which $M_i \succ 0$ for all i, and we choose an arbitrary pair of primal and dual optimal solutions $\Big((X^*, Y^*, \boldsymbol{\lambda}^*), \boldsymbol{\mu}^*\Big)$. The dual feasibility relation implies $\boldsymbol{\mu}^* \neq \mathbf{0}$, and so $\sum_i \mu_i^* M_i$ is a positive definite matrix (we exclude the trivial case $C = 0$). Since C is of rank r, we deduce that

$$\operatorname{rank}(\sum_i \mu_i^* M_i - C) \geq n - r.$$

Finally, the complementary slackness relation indicates that the columns of X^* belong to the nullspace of $(\sum_i \mu_i^* M_i - C)$, which is a vector space of dimension at most $n - (n - r) = r$, and so we conclude that $\operatorname{rank} X^* \leq r$.

We now turn to the study of the general case in which $M_i \succeq 0$. To this end, we consider the perturbed problems

$$\begin{aligned}
\max \quad & \langle C, X \rangle + \langle R_0, Y \rangle + \boldsymbol{h_0}^T \boldsymbol{\lambda} \\
\text{s.t.} \quad & \langle M_i + \varepsilon \boldsymbol{I}, X \rangle \leq b_i + \langle R_i, Y \rangle + \boldsymbol{h_i}^T \boldsymbol{\lambda} \qquad i \in [s], \qquad (P^{\varepsilon}_{\text{CMB}}) \\
& X \succeq 0, \ Y \succeq 0,
\end{aligned}$$

and

$$\begin{aligned}
\min_{\boldsymbol{\mu} \geq 0} \quad & \sum_{i=1}^{s} \mu_i b_i, \\
\text{s.t.} \quad & \sum_{i=1}^{s} \mu_i (M_i + \varepsilon \boldsymbol{I}) \succeq C, \qquad (D^{\varepsilon}_{\text{CMB}}) \\
& R_0 + \sum_{i=1}^{s} \mu_i R_i \preceq 0, \\
& \boldsymbol{h_0} + H\boldsymbol{\mu} = \mathbf{0}.
\end{aligned}$$

where $\varepsilon \geq 0$. Note that the refined Slater constraints qualification for the unperturbed problems (P_{CMB}) and (D_{CMB}) (i.e. simultaneous feasibility (resp. strict feasibility) of all

the affine constraints (resp. non-affine constraints)) implies the qualification of the constraints for $(P^{\varepsilon}_{\text{CMB}})$ and $(D^{\varepsilon}_{\text{CMB}})$ on a neighborhood $\varepsilon \in [0, \varepsilon_0]$, $\varepsilon_0 > 0$. We denote by $\Big((X^{\varepsilon}, Y^{\varepsilon}, \boldsymbol{\lambda}^{\boldsymbol{\varepsilon}}), \boldsymbol{\mu}^{\boldsymbol{\varepsilon}}\Big)$ a pair of primal and dual solutions of $(P^{\varepsilon}_{\text{CMB}})$–$(D^{\varepsilon}_{\text{CMB}})$. If $\varepsilon > 0$, $M_i + \varepsilon \boldsymbol{I} \succ 0$ and it follows from the previous discussion that X^{ε} is of rank at most r. We show below that we can choose the optimal variables $(X^{\varepsilon}, Y^{\varepsilon}, \boldsymbol{\lambda}^{\boldsymbol{\varepsilon}}, \boldsymbol{\mu}^{\boldsymbol{\varepsilon}})_{\varepsilon \in]0, \varepsilon_0]}$ within a bounded region, so that we can construct a converging subsequence $(X^{\varepsilon_k}, Y^{\varepsilon_k}, \boldsymbol{\lambda}^{\boldsymbol{\varepsilon_k}}, \boldsymbol{\mu}^{\boldsymbol{\varepsilon_k}})_{k \in \mathbb{N}}$, $\varepsilon_k \to 0$ from these variables. To conclude, we will see that the limit $(X^0, Y^0, \boldsymbol{\lambda^0}, \boldsymbol{\mu^0})$ satisfies the KKT conditions for Problems (P_{CMB})–(D_{CMB}), and that X^0 is of rank at most r.

Let us denote the optimal value of Problems $(P^{\varepsilon}_{\text{CMB}})$–$(D^{\varepsilon}_{\text{CMB}})$ by $OPT(\varepsilon)$. Since the constraints of the primal problem becomes tighter when ε grows, it is clear that $OPT(\varepsilon)$ is nonincreasing with respect to ε, so that

$$\forall \varepsilon \in [0, \varepsilon_0],\ OPT(\varepsilon_0) \leq OPT(\varepsilon) \leq OPT(0).$$

Now let $\varepsilon \in]0, \varepsilon_0]$. By assumption, there exists a vector $\overline{\boldsymbol{\mu}} \geq \mathbf{0}$ such that

$$\sum_i \overline{\mu}_i (M_i + \varepsilon \boldsymbol{I}) \succeq \sum_i \overline{\mu}_i M_i \succ C, \quad \text{and} \quad R_0 + \sum_i \overline{\mu}_i R_0 \prec 0. \tag{4.16}$$

Therefore, we have

$$\begin{aligned} OPT(\varepsilon) &= \langle C, X^{\varepsilon} \rangle + \langle R_0, Y^{\varepsilon} \rangle + \boldsymbol{h_0}^T \boldsymbol{\lambda}^{\boldsymbol{\varepsilon}} \\ &\leq \Big\langle \sum_i \overline{\mu}_i (M_i + \varepsilon \boldsymbol{I}), X^{\varepsilon} \Big\rangle + \langle R_0, Y^{\varepsilon} \rangle + \boldsymbol{h_0}^T \boldsymbol{\lambda}^{\boldsymbol{\varepsilon}} \\ &\leq \sum_i \overline{\mu}_i \Big(b_i + \langle R_i, Y^{\varepsilon} \rangle + \boldsymbol{h_i}^T \boldsymbol{\lambda}^{\boldsymbol{\varepsilon}} \Big) + \langle R_0, Y^{\varepsilon} \rangle + \boldsymbol{h_0}^T \boldsymbol{\lambda}^{\boldsymbol{\varepsilon}} \\ &= \overline{\boldsymbol{\mu}}^T \boldsymbol{b} + \langle \sum_i \overline{\mu}_i R_i + R_0, Y^{\varepsilon} \rangle + (\underbrace{\boldsymbol{h_0} + H \overline{\boldsymbol{\mu}}}_{=\mathbf{0}})^T \boldsymbol{\lambda}^{\boldsymbol{\varepsilon}}, \end{aligned}$$

where the first inequality follows from (4.16), and the second one from the feasibility condition $\langle M_i + \varepsilon \boldsymbol{I}, X^{\varepsilon} \rangle \leq b_i + \langle R_i, Y^{\varepsilon} \rangle + \boldsymbol{h_i}^T \boldsymbol{\lambda}^{\boldsymbol{\varepsilon}}$. The assumption (4.16) moreover implies that $-(\sum_i \overline{\mu}_i R_i + R_0)$ is positive definite, so that its smallest eigenvalue λ' is positive, and

$$\lambda' \ \text{trace}\ Y^{\varepsilon} \leq \Big\langle -(\sum_i \overline{\mu}_i R_i + R_0), Y^{\varepsilon} \Big\rangle \leq \overline{\boldsymbol{\mu}}^T \boldsymbol{b} - OPT(\varepsilon) \leq \overline{\boldsymbol{\mu}}^T \boldsymbol{b} - OPT(\varepsilon_0).$$

This shows that the trace of Y^{ε} is bounded, and so $Y^{\varepsilon} \succeq 0$ is bounded.

Similarly, to bound X^ε, we write:

$$\begin{aligned}\langle \sum_i \overline{\mu}_i M_i - C, X^\varepsilon \rangle &\leq \langle \sum_i \overline{\mu}_i (M_i + \varepsilon \boldsymbol{I}) - C, X^\varepsilon \rangle \\ &= \langle \sum_i \overline{\mu}_i (M_i + \varepsilon \boldsymbol{I}), X^\varepsilon \rangle - OPT(\varepsilon) + \langle R_0, Y^\varepsilon \rangle + \boldsymbol{h_0}^T \boldsymbol{\lambda}^\varepsilon \\ &\leq \sum_i \overline{\mu}_i \big(b_i + \langle R_i, Y^\varepsilon \rangle + \boldsymbol{h_i}^T \boldsymbol{\lambda}^\varepsilon \big) - OPT(\varepsilon) + \langle R_0, Y^\varepsilon \rangle + \boldsymbol{h_0}^T \boldsymbol{\lambda}^\varepsilon \\ &= \overline{\boldsymbol{\mu}}^T \boldsymbol{b} - OPT(\varepsilon) + \underbrace{\langle \sum_i \overline{\mu}_i R_i + R_0, Y^\varepsilon \rangle}_{\leq 0} + \underbrace{(\boldsymbol{h_0} + H\overline{\boldsymbol{\mu}})}_{=\boldsymbol{0}}{}^T \boldsymbol{\lambda}^\varepsilon,\end{aligned}$$

where the first equality comes from the expression of $OPT(\varepsilon)$. The matrix $\sum_i \overline{\mu}_i M_i - C$ is positive definite and its smallest eigenvalue λ'' is therefore positive. Hence,

$$\lambda'' \ \text{trace} \ X^\varepsilon \leq \overline{\boldsymbol{\mu}}^T \boldsymbol{b} - OPT(\varepsilon) \leq \overline{\boldsymbol{\mu}}^T \boldsymbol{b} - OPT(\varepsilon_0),$$

and this shows that the matrix $X^\varepsilon \succeq 0$ is bounded.

Now, note that the feasibility of $\boldsymbol{\lambda}^\varepsilon$ implies that the quantity $b_i + \langle R_i, Y^\varepsilon \rangle + \boldsymbol{h_i}^T \boldsymbol{\lambda}^\varepsilon$ is nonnegative for all $i \in [s]$. Since Y^ε is bounded, we deduce the existence of a lower bound $m_i \in \mathbb{R}$ such that $\boldsymbol{h_i}^T \boldsymbol{\lambda}^\varepsilon \geq m_i$ $(\forall i \in [s])$. Similarly, since $\boldsymbol{h_0}^T \boldsymbol{\lambda}^\varepsilon \geq OPT(\varepsilon_0) - \langle C, X^\varepsilon \rangle - \langle R_0, Y^\varepsilon \rangle$, there is a scalar m_0 such that $\boldsymbol{h_0}^T \boldsymbol{\lambda}^\varepsilon \geq m_0$. We now use the fact that every vector $(-\boldsymbol{h_i})$ may be written as a positive combination of the $\boldsymbol{h_k}$, $(k \in \{0\} \cup [s])$, and we obtain that the quantities $\boldsymbol{h_i}^T \boldsymbol{\lambda}^\varepsilon$ are also bounded from above. Let us denote by H_0 the matrix $[\boldsymbol{h_0}, H]$; we have just proved that the vector $H_0^T \boldsymbol{\lambda}^\varepsilon$ is bounded:

$$\exists \overline{m} \in \mathbb{R} : \quad \| H_0^T \boldsymbol{\lambda}^\varepsilon \|_2 \leq \overline{m}$$

(the latter bound does not depend on ε). Note that one may assume without loss of generality that $\boldsymbol{\lambda}^\varepsilon \in \operatorname{Im} H_0$ (otherwise we consider the projection $\boldsymbol{\lambda}^\varepsilon_{\boldsymbol{P}}$ of $\boldsymbol{\lambda}^\varepsilon$ on $\operatorname{Im} H_0$ which is also a solution since $H_0^T \boldsymbol{\lambda}^\varepsilon = H_0^T \boldsymbol{\lambda}^\varepsilon_{\boldsymbol{P}}$. We know from the Courant-Fisher theorem that the smallest positive eigenvalue of $H_0 H_0^T$ satisfies:

$$\lambda^>_{\min}(H_0 H_0^T) = \min_{\boldsymbol{v} \in \operatorname{Im} H_0 \setminus \{\boldsymbol{0}\}} \frac{\boldsymbol{v}^T H_0 H_0^T \boldsymbol{v}}{\boldsymbol{v}^T \boldsymbol{v}}.$$

Therefore, since we have assumed $\boldsymbol{\lambda}^\varepsilon \in \operatorname{Im} H_0$:

$$\| \boldsymbol{\lambda}^\varepsilon \|^2 \leq \frac{\| H_0^T \boldsymbol{\lambda}^\varepsilon \|^2}{\lambda^>_{\min}(H_0 H_0^T)} \leq \frac{\overline{m}^2}{\lambda^>_{\min}(H_0 H_0^T)}.$$

It remains to show that the dual optimal variable $\boldsymbol{\mu}^\varepsilon$ is bounded. Our strict primal feasibility assumption (which does not entail generality thanks to Lemma 4.3.1) ensures the existence of a matrix $\overline{Y} \succ 0$ and a vector $\overline{\boldsymbol{\lambda}}$ such that

$$\forall i \in [s], \ \langle R_i, \overline{Y} \rangle + b_i + \boldsymbol{h_i}^T \overline{\boldsymbol{\lambda}} = \eta_i > 0.$$

By dual feasibility, $R_0 + \sum_i \mu_i^\varepsilon R_i$ is a negative semidefinite matrix, and we have:

$$0 \geq \langle R_0, \overline{Y} \rangle + \sum_{i=1}^{s} \mu_i^\varepsilon \langle R_i, \overline{Y} \rangle = \langle R_0, \overline{Y} \rangle + \sum_{i=1}^{s} \mu_i^\varepsilon (\eta_i - b_i - \boldsymbol{h_i}^T \overline{\boldsymbol{\lambda}}).$$

Hence, we have the following inequalities:

$$\begin{aligned} \forall k \in [s],\ \eta_k \mu_k^\varepsilon \leq \sum_{i=1}^{s} \eta_i \mu_i^\varepsilon &\leq \boldsymbol{b}^T \boldsymbol{\mu}^{\boldsymbol{\varepsilon}} + \overline{\boldsymbol{\lambda}}^T H \boldsymbol{\mu}^{\boldsymbol{\varepsilon}} - \langle R_0, \overline{Y} \rangle \\ &= OPT(\varepsilon) - \overline{\boldsymbol{\lambda}}^T \boldsymbol{h_0} - \langle R_0, \overline{Y} \rangle \\ &\leq OPT(0) - \overline{\boldsymbol{\lambda}}^T \boldsymbol{h_0} - \langle R_0, \overline{Y} \rangle, \end{aligned}$$

and we have shown that $\boldsymbol{\mu}^{\boldsymbol{\varepsilon}} \geq \mathbf{0}$ is bounded.

We can therefore construct a sequence of pairs of primal and dual optimal solutions $(X^{\varepsilon_k}, Y^{\varepsilon_k}, \boldsymbol{\lambda}^{\boldsymbol{\varepsilon_k}}, \boldsymbol{\mu}^{\boldsymbol{\varepsilon_k}})_{k \in \mathbb{N}}$ that converges, with $\varepsilon_k \underset{k \to \infty}{\longrightarrow} 0$, $\varepsilon_k > 0$. In this sequence, the limit X^0 of X^{ε_k} is of rank at most r, because the rank is a lower semicontinuous function and $\text{rank } X^{\varepsilon_k} \leq r$ for all $k \in \mathbb{N}$. It remains to show that $(X^0, Y^0, \boldsymbol{\lambda^0})$ is a solution of Problem (P_{CMB}). The $\varepsilon-$perturbed KKT conditions must hold for all $k \in \mathbb{N}$, and so they hold for the pair $\left((X^0, Y^0, \boldsymbol{\lambda^0}), \boldsymbol{\mu^0}\right)$ by taking the limit (this works because $\mathbb{S}_m^+$ is closed). This concludes the proof of the existence of a solution in which $\text{rank } X \leq r$.

It remains to show the second statement of this theorem, namely that if $C \neq 0$ and $\overline{r} := \min_{i \in [s]} \text{rank } M_i$, then the rank of X is bounded by $n - \overline{r} + r$ for any solution $(X, Y, \boldsymbol{\lambda})$ of (P_{CMB}).

Let $(X^*, Y^*, \boldsymbol{\lambda}^*)$ be a solution of Problem (P_{CMB}). If the primal problem is strictly feasible, then there exists a Lagrange multiplier $\boldsymbol{\mu}^* \geq \mathbf{0}$ such that the KKT conditions described at the beginning of this proof are satisfied. Since $C \neq 0$, we have $\boldsymbol{\mu}^* \neq \mathbf{0}$, and we can write:

$$\text{rank } (\sum_{i \in [s]} \mu_i^* M_i - C) \geq \overline{r} - r.$$

Hence, since by complementary slackness, X^* belongs to the nullspace of $(\sum_{i \in [s]} \mu_i^* M_i - C)$, we find $\text{rank } X^* \leq n - \overline{r} + r$.

If the primal problem is not strictly feasible, there must be an index $i \in [s]$ such that $\langle M_i, X^* \rangle = 0$ (otherwise, $(\eta_1 \boldsymbol{I}, Y^* + \eta_2 \boldsymbol{I}, \boldsymbol{\lambda}^*)$ would be strictly feasible for sufficiently small positive reals η_1 and η_2). Therefore, X^* is in the nullspace of a matrix of rank larger than $\overline{r}$, and $\text{rank } X^* \leq n - \overline{r} \leq n - \overline{r} + r$. □

4.3.3 Proof of Theorem 4.2.1

We assume that Problems (P_{CMB}) and (D_{CMB}) are feasible, and for $\eta \geq 0$ we consider the following pair of primal and dual perturbed problems.

$$\begin{array}{ll} \sup & \langle C, X\rangle + \langle R_0, Y\rangle + \boldsymbol{h_0}^T \boldsymbol{\lambda} \\ \text{s.t.} & \langle M_i, X\rangle \leq b_i + \langle R_i, Y\rangle + \boldsymbol{h_i}^T \boldsymbol{\lambda} \qquad i \in [s], \\ & \eta\ (\text{trace}\ X + \text{trace}\ Y) \leq 1, \\ & X \succeq 0,\ Y \succeq 0, \end{array} \tag{P_η}$$

and

$$\begin{array}{lll} \inf\limits_{\boldsymbol{\mu}\geq \mathbf{0},\ \sigma \geq 0} & \sum\limits_{i=1}^{s} \mu_i b_i + \sigma, & \\ \text{s.t.} & \sum\limits_{i=1}^{s} \mu_i M_i + \sigma\eta \boldsymbol{I} \succeq C, & (D_\eta) \\ & R_0 + \sum\limits_{i=1}^{s} \mu_i R_i - \sigma\eta \boldsymbol{I} \preceq 0, & \\ & \boldsymbol{h_0} + H\boldsymbol{\mu} = \mathbf{0}. & \end{array}$$

It is clear that the feasibility of Problem (P_{CMB}) implies that of (P_η) if $\eta > 0$ is sufficiently small. Let $\overline{\boldsymbol{\mu}}$ be a dual feasible variable for Problem (D_{CMB}), and $\sigma > 0$ be sufficiently large so that $\sum_{i=1}^{s} \overline{\mu}_i M_i + \sigma\eta \boldsymbol{I} \succ C$ and $R_0 + \sum_{i=1}^{s} \overline{\mu}_i R_i - \sigma\eta \boldsymbol{I} \prec 0$: the refined Slater condition holds for the perturbed problem (D_η). Hence, by Theorem 4.2.2, there exists a solution $(X^\eta, Y^\eta, \boldsymbol{\lambda^\eta})$ of Problem (P_η) in which $\text{rank}\ X^\eta \leq r$. We next show that $\langle C, X^\eta\rangle + \langle R_0, Y^\eta\rangle + \boldsymbol{h_0}^T \boldsymbol{\lambda^\eta}$ converges to the value of the supremum in Problem (P_{CMB}) as $\eta \to 0^+$, which will complete this proof.

Let η_k be a positive sequence decreasing to 0, and define $\gamma_k := \langle C, X^{\eta_k}\rangle + \langle R_0, Y^{\eta_k}\rangle + \boldsymbol{h_0}^T \boldsymbol{\lambda^{\eta_k}}$. It is clear that γ_k is a nondecreasing sequence, because the constraints in Problem (P_η) become looser as η gets smaller, and γ_k is bounded from above by the value of the supremum γ^* in Problem (P_{CMB}). Therefore, $(\gamma_k)_{k\in\mathbb{N}}$ converges. Assume (*ad absurdum*) that the limit of this sequence is $\gamma_\infty < \gamma*$. Then, there are some variables $(X_0, Y_0, \boldsymbol{\lambda_0})$ that are feasible for (P_{CMB}), and such that $\langle C, X_0\rangle + \langle R_0, Y_0\rangle + \boldsymbol{h_0}^T \boldsymbol{\lambda_0} > \gamma_\infty$. But then, $(X_0, Y_0, \boldsymbol{\lambda_0})$ is also feasible for Problem (P_η), when $\eta \leq \eta_0 := (\text{trace}\ X_0 + \text{trace}\ Y_0)^{-1}$. For any $k \in \mathbb{N}$ such that $\eta_k \leq \eta_0$, this contradicts the optimality of $(X^{\eta_k}, Y^{\eta_k}, \boldsymbol{\lambda^{\eta_k}})$ for Problem (P_{η_k}). Hence, $\gamma_\infty = \gamma*$ and the proof is complete.

Chapter 5

The Second Order Cone Programming approach

This chapter essentially recalls the results of [Sag09b]. We shall see that many optimal experimental design problems can be formulated as *Second order cone programs* (SOCP). Unlike the SDP formulations of Chapter 3, the SOCP arising in optimal experimental design remain tractable on very large instances. In addition, the second order cone programming is a convenient framework which offers both modelling flexibility and theoretical safeguards.

The proposed second order cone programming approach arises naturally from a geometrical characterization of $\boldsymbol{c}-$optimality for multiresponse experiments. However, this geometric point of view leaves unexplained the equivalence between the formerly known SDPs (cf. Section 3.3) and the new SOCPs. In fact, most results from this chapter admit an alternative proof relying on the rank reduction theorem of Chapter 4.

5.1 An Elfving Theorem for multiresponse experiments

In this section, we extend the result of Elfving (Theorem 2.4.1) to the case of multidimensional observations. For the sake of generality, we turn temporarily back to the general case in which the regression region $\mathcal{X}$ is a (possibly infinite) compact set. Throughout this section, we will also make the assumption that every observation is of dimension l (i.e. $l(\boldsymbol{x}) = l$ for all $\boldsymbol{x} \in \mathcal{X}$). We point out that this assumption is made with the only goal to simplify the notation, and does not entail the generality (we handle the case in which the experiment at $\boldsymbol{x}$ only gives $k < l$ measurements by setting $l - k$ rows of the matrix $A(\boldsymbol{x})$ to zero).

Some results of this chapter, including Theorem 5.1.1, were presented at the conference [SBG09], and the technical result justifying the reduction to a SOCP was posted on arXiv [Sag09a]. Shortly before the time of submission, Dette and Holland-Letz published an article in *Annals of Statistics*, in which Theorem 5.1.1 was established independently

(Theorem 3.3 in [DHL09]). They considered a heteroscedastic model (i.e. an experimental model where both the mean and the variance of the observations depend on the parameter of interest), which led them to study the case in which the observation matrices are of rank $k \geq 2$, just as in the model of *multiresponse experiments*. They used their geometrical characterization of the $\boldsymbol{c}-$optimal design for heteroscedastic models in an application to toxicokinetics and pharmacokinetics. It should also be mentioned that the proof of Dette and Holland-Letz relies on an equivalence theorem (Theorem 3.1 in [DHL09]), while ours is closer to Elfving's original approach, as done previously by Studden [Stu05] for other results in optimal design of experiments. The main result of our article (reduction to a SOCP, Theorem 5.2.1), provides a new insight on the relations between these two approaches : they are actually dual from each other (in the Lagrangian sense). Indeed, the approach of Dette and Holland-Letz corresponds to the optimality conditions of the primal SOCP (5.3), while our direct geometrical characterization corresponds to the dual SOCP (5.4), and strong duality holds between these two optimization problems.

5.1.1 c-optimality

To state our result, we will need the following generalization of the Elfving set 2.20 for multiresponse experiments:

$$\overline{\mathcal{E}} = \operatorname{conv}\Big(\{A(\boldsymbol{x})^T\boldsymbol{\epsilon},\ \ \boldsymbol{x} \in \mathcal{X},\ \boldsymbol{\epsilon} \in \mathbb{R}^l,\ \|\boldsymbol{\epsilon}\|_2 \leq 1\}\Big).$$

Note that $\overline{\mathcal{E}}$ is a generalization of the classical Elfving set (the factor ± 1 has been substituted by a vector $\boldsymbol{\epsilon}$ in the unit ball of $\mathbb{R}^l$).

Theorem 5.1.1 (Extension of Elfving's theorem to the case of multiresponse experiments)**.** *A design $\xi = \{\boldsymbol{x_i}, w_i\}$ is $\boldsymbol{c}-$optimal if and only if there exists a positive scalar t and vectors $\boldsymbol{\epsilon_i}$ in the unit ball of $\mathbb{R}^l$ (i.e. $\|\boldsymbol{\epsilon_i}\|_2 \leq 1$), such that*

$$t\boldsymbol{c} = \sum_i w_i A(\boldsymbol{x_i})^T\boldsymbol{\epsilon_i} \in \partial\overline{\mathcal{E}}.$$

Moreover, $t^{-2} = \boldsymbol{c}^T M(\xi)^{-}\boldsymbol{c}$ is the minimal variance.

Proof. We consider an unbiased linear estimator for $\zeta = \boldsymbol{c}^T\boldsymbol{\theta}$:

$$\hat{\zeta} = \boldsymbol{h}^T\boldsymbol{y}(\xi),\ \text{with}\ \boldsymbol{h} = [\boldsymbol{h_1}^T, ..., \boldsymbol{h_s}^T]^T \in \mathbb{R}^{sl}, \quad \boldsymbol{h_i} \in \mathbb{R}^l.$$

The unbiasedness property forces the following equality to hold :

$$A(\xi)^T\boldsymbol{h} = \sum_{i=1}^{s} A(\boldsymbol{x_i})^T\boldsymbol{h_i} = \boldsymbol{c}.$$

Now, the Cauchy-Schwarz inequality gives the following lower bound for the variance of $\hat{\zeta}$:

$$\mathrm{Var}(\hat{\zeta}) = \boldsymbol{h}^T \Delta(\boldsymbol{w}) \boldsymbol{h} = \sum_{k=1}^{s} \frac{\|\boldsymbol{h_k}\|^2}{w_k} \geq \Big(\sum_{k=1}^{s} \|\boldsymbol{h_k}\| \Big)^2, \tag{5.1}$$

where $\| \cdot \|$ denotes the L_2 norm and $\Delta(\boldsymbol{w})$ was defined in Equation (2.5). We recall that we assume $\boldsymbol{w} > \boldsymbol{0}$ without loss of generality, since an experiment with a zero weight can be removed from the design ξ.

We show that $\frac{\boldsymbol{c}}{\sum_k \|\boldsymbol{h_k}\|} \in \overline{\mathcal{E}}$, by writing:

$$\frac{\boldsymbol{c}}{\sum_k \|\boldsymbol{h_k}\|} = \frac{A(\xi)^T \boldsymbol{h}}{\sum_k \|\boldsymbol{h_k}\|} = \sum_i A(\boldsymbol{x_i})^T \frac{\boldsymbol{h_i}}{\sum_k \|\boldsymbol{h_k}\|} = \sum_{\{i:\|\boldsymbol{h_i}\|>0\}} \mu_i A(\boldsymbol{x_i})^T \epsilon_i,$$

where $\mu_i = \frac{\|\boldsymbol{h_i}\|}{\sum_k \|\boldsymbol{h_k}\|}$ and $\epsilon_i = \frac{\boldsymbol{h_i}}{\|\boldsymbol{h_i}\|}$, so that $\|\boldsymbol{\epsilon_i}\| = 1$, $\mu_i \geq 0$ and $\sum_i \mu_i = 1$.

Let t be a positive scalar such that $t\boldsymbol{c} \in \partial\overline{\mathcal{E}}$. The fact that $\frac{\boldsymbol{c}}{\sum_k \|\boldsymbol{h_k}\|} \in \overline{\mathcal{E}}$ implies

$$\frac{1}{\sum_k \|\boldsymbol{h_k}\|} \leq t \Longrightarrow \Big(\sum_{k=1}^{s} \|\boldsymbol{h_k}\| \Big)^2 \geq t^{-2}. \tag{5.2}$$

Combining (5.1) and (5.2) leads to the lower bound t^{-2} for the variance of any linear unbiased estimator of ζ.

We will show that this lower bound is attained if and only if the design ξ satisfies the condition of the theorem. To do this, notice that for a design ξ and an estimator $\boldsymbol{h}^T \boldsymbol{y}(\xi)$ to be optimal, it is necessary and sufficient that the inequalities (5.1) and (5.2) are equalities. The Cauchy-Schwarz inequality (5.1) is an equality if and only if $\boldsymbol{w}$ is proportional to the vector $[\|\boldsymbol{h_1}\|, ..., \|\boldsymbol{h_s}\|]^T$, i.e.

$$w_i = \frac{\|\boldsymbol{h_i}\|}{\sum_k \|\boldsymbol{h_k}\|}.$$

The second inequality (5.2) is an equality whenever $\frac{\boldsymbol{c}}{\sum_k \|\boldsymbol{h_k}\|} \in \partial\overline{\mathcal{E}}$, i.e. $\frac{1}{\sum_k \|\boldsymbol{h_k}\|} = t$, where t is the largest real such that $t\boldsymbol{c} \in \overline{\mathcal{E}}$. We can write

$$\partial\overline{\mathcal{E}} \ni t\boldsymbol{c} = t \sum_i A(\boldsymbol{x_i})^T \boldsymbol{h_i} = \sum_{\{i:\|\boldsymbol{h_i}\|>0\}} \mu_i A(\boldsymbol{x_i})^T \boldsymbol{\epsilon_i},$$

with $\mu_i = t\|\boldsymbol{h_i}\|$ and $\boldsymbol{\epsilon_i} = \frac{\boldsymbol{h_i}}{\|\boldsymbol{h_i}\|}$. We have $\|\boldsymbol{\epsilon_i}\| = 1$, and the equality conditions are satisfied if and only if $\mu_i = w_i$. □

As a consequence of this theorem, we will see In Section 5.2.1 that the $\boldsymbol{c}-$optimal design of finitely many multiresponse experiments can be formulated as a second order cone program (SOCP).

5.1.2 The case of A-optimality

When there are several quantities of interest, i.e. when $\boldsymbol{\zeta}$ consists in a collection of r linear combinations of the parameters ($\boldsymbol{\zeta} = K^T\boldsymbol{\theta}$, where $K = [\boldsymbol{c_1}, \ldots, \boldsymbol{c_r}]$ is $m \times r$), the $A-$optimal problem is to find the design ξ that minimizes $\text{trace}\,(K^T(M(\xi))^- K)$. We recall that an interesting case occurs when $K = I$, i.e. when the experimenter wants to estimate the whole vector of parameters (cf. Section 2.3.2).

We show in this section that computing the $A-$optimal design for $K^T\boldsymbol{\theta}$ can be written as a $\boldsymbol{c}-$optimal design problem with multidimensional observations. Up to the factor $\frac{1}{m}$, the objective function of (2.18) can indeed be written as

$$\text{trace}\,(K^T M(\xi)^- K) = \sum_{k=1}^{r} \boldsymbol{c_k}^T M(\xi)^- \boldsymbol{c_k}.$$

We now define the vector $\tilde{\boldsymbol{c}}$ as the vertical concatenation of the columns $\boldsymbol{c_i}$, i.e. $\tilde{\boldsymbol{c}} = [\boldsymbol{c_1}^T, ..., \boldsymbol{c_r}^T]^T$. Now , we have: $\text{trace}(K^T M(\xi)^- K) = \tilde{\boldsymbol{c}}^T \tilde{M}(\xi)^- \tilde{\boldsymbol{c}}$, where:

$$\begin{aligned}
\tilde{M}(\xi) = \begin{pmatrix} M(\xi) & & \\ & \ddots & \\ & & M(\xi) \end{pmatrix} &= \sum_{i=1}^{s} w_i \begin{pmatrix} A(\boldsymbol{x_i})^T A(\boldsymbol{x_i}) & & \\ & \ddots & \\ & & A(\boldsymbol{x_i})^T A(\boldsymbol{x_i}) \end{pmatrix} \\
&= \sum_{i=1}^{s} w_i \begin{pmatrix} A(\boldsymbol{x_i}) & & \\ & \ddots & \\ & & A(\boldsymbol{x_i}) \end{pmatrix}^T \overbrace{\begin{pmatrix} A(\boldsymbol{x_i}) & & \\ & \ddots & \\ & & A(\boldsymbol{x_i}) \end{pmatrix}}^{\tilde{A}(\boldsymbol{x_i})} \\
&= \sum_{i=1}^{s} w_i \tilde{A}(\boldsymbol{x_i})^T \tilde{A}(\boldsymbol{x_i}).
\end{aligned}$$

In the latter equation, $\tilde{A}(\boldsymbol{x_i})$ contains r blocks and is of dimension $rl \times rm$. We can now rewrite Problem (2.18) in the following form:

$$\begin{aligned}
&\min_{\xi} \ \text{trace}\,(\tilde{\boldsymbol{c}}^T \tilde{M}(\xi)^- \tilde{\boldsymbol{c}}) \\
&\text{s.t.} \ \sum_{i=1}^{s} w_i = 1 \\
&\qquad \tilde{M}(\xi) = \sum_{i=1}^{s} w_i \tilde{A}(\boldsymbol{x_i})^T \tilde{A}(\boldsymbol{x_i}) \\
&\qquad \forall\ i \in [s], w_i \geq 0, \boldsymbol{x_i} \in \mathcal{X}.
\end{aligned}$$

We have thus shown that the problem of finding the $A-$optimal design is nothing but a $\tilde{\boldsymbol{c}}-$optimal design problem, with augmented observation matrices $\tilde{A}(\boldsymbol{x_i})$. As a consequence, our result of reduction of the $\boldsymbol{c}-$optimal design problem (Section 5.2.1) also applies for the more general class of $A-$optimal design problem for a subsystem $K^T\boldsymbol{\theta}$ of the parameters

(cf. Section 5.2.2).

We now show that the geometrical characterization in Theorem 5.1.1 generalizes the result of Studden [Stu71], who established an Elfving type result for the characterization of $A-$optimal designs in the case of scalar observations ($l = 1$ and $A(\boldsymbol{x}) = \boldsymbol{a_x}^T$ is a row vector). This characterization is based on the following extension of the Elfving set when the matrix K is $m \times r$:

$$\mathcal{E}_S = \text{conv}\left(\{\boldsymbol{a_x}\boldsymbol{\epsilon}^T | \boldsymbol{x} \in \mathcal{X},\ \boldsymbol{\epsilon} \in \mathbb{R}^r,\ \|\boldsymbol{\epsilon}\| \leq 1\}\right) \subset \mathbb{R}^{m \times r}$$

Theorem 5.1.2 (Studden,1971). *A design $\xi = \{\boldsymbol{x_i}, w_i\}$ is $A-$optimal for $K^T\boldsymbol{\theta}$ if and only if there exists a scalar $t > 0$ and vectors $\boldsymbol{\epsilon_i}$ in the unit ball of $\mathbb{R}^r$ such that*

$$tK = \sum_i w_i \boldsymbol{a_{x_i}}\boldsymbol{\epsilon_i}^T \in \partial\mathcal{E}_S.$$

Moreover, $t^{-2} = \text{trace}(K^T M(\xi)^- K)$ is the optimal value of the $A-$criterion.

One can easily verify that this theorem is a particular case of Theorem 5.1.1. Using the previously introduced notation indeed, Theorem 5.1.1 says that $\xi = \{\boldsymbol{x_i}, w_i\}$ is $A-$optimal for $K^T\boldsymbol{\theta}$ if and only if there exists a scalar $t > 0$ and vectors $\boldsymbol{\epsilon_i}$ in the unit ball of $\mathbb{R}^{rl}$ such that

$$t\tilde{\boldsymbol{c}} = \sum_i w_i \tilde{A}(\boldsymbol{x_i})^T \boldsymbol{\epsilon_i} \in \partial\overline{\mathcal{E}},$$

and we notice that $\tilde{\boldsymbol{c}}$ is the vectorized version of K, and when $l = 1$, $\overline{\mathcal{E}}$ is the vectorized version of $\mathcal{E}_S$ and $\tilde{A}(\boldsymbol{x_i})^T\boldsymbol{\epsilon_i} = [\epsilon_{i1}\boldsymbol{a_{x_i}}^T, \ldots, \epsilon_{is}\boldsymbol{a_{x_i}}^T]^T$ is the vectorized version of $\boldsymbol{a_{x_i}}\boldsymbol{\epsilon_i}^T$.

5.2 The Second order cone programming approach

In this section, we will see that many optimal design problems can be formulated as Second Order Cone Programs when the regression region is finite, i.e. $\mathcal{X} = [s]$. We come back to the initial notation, where l_i denotes the first dimension of the observation matrix A_i (we do not assume $l_i = l$ for all i anymore).

5.2.1 c-optimality

We show in this section that the $\boldsymbol{c}-$optimal design problem reduces to a Second Order Cone Program (SOCP). We will give two proofs of this result : the first one is a consequence of our generalization of the Elfving theorem to the case of multiresponse experiments (Theorem 5.1.1). The second proof uses the rank reduction theorem of Chapter 4.

Theorem 5.2.1 (Computation of the $c-$optimal design by SOCP). *Let $\boldsymbol{u}^*, (\boldsymbol{\mu}^*, \boldsymbol{h_i^*})$ be a pair of primal and dual solutions of the second order cone programs:*

$$
\begin{aligned}
\textit{(P-SOCP)}: \quad & \max_{\boldsymbol{u}\in\mathbb{R}^m} \quad \boldsymbol{c}^T\boldsymbol{u} && (5.3)\\
& \qquad \forall i \in [s], \quad \|A_i\boldsymbol{u}\| \leq 1 \\
\textit{(D-SOCP)}: \quad & \min_{\boldsymbol{\mu}\in\mathbb{R}^s,\ \boldsymbol{h_i}\in\mathbb{R}^{l_i}} \quad \sum_i \mu_i && (5.4)\\
& \qquad \boldsymbol{c} = \sum_i A_i^T \boldsymbol{h_i} \\
& \qquad \forall i \in [s], \quad \|\boldsymbol{h_i}\| \leq \mu_i.
\end{aligned}
$$

We define

$$\boldsymbol{w} := t\boldsymbol{\mu}^*, \qquad where \qquad t = (\sum_{k=1}^{s} \mu_k^*)^{-1}.$$

Then $\boldsymbol{w}$ is a $\boldsymbol{c}-$optimal design. Moreover, $\hat{\zeta} = \sum \boldsymbol{h_i^*}^T \boldsymbol{y_i}$ is the best linear estimator of $\boldsymbol{c}^T\boldsymbol{\theta}$, and the optimal variance is $\mathrm{var}(\hat{\zeta}) = t^{-2} = (\sum_i \mu_i^*)^2 = (\boldsymbol{c}^T\boldsymbol{u}^*)^2$.

Proof relying on the extended Elfving theorem

Proof. This result is actually a corollary of Theorem 5.1.1. As in the proof of the latter theorem, define t as the largest scalar such that $t\boldsymbol{c} \in \overline{\mathcal{E}}$, i.e. such that there exists w_i summing to 1 and vectors $\boldsymbol{\epsilon_i}$ in the unit ball of $\mathbb{R}^l$ satisfying

$$t\boldsymbol{c} = \sum_{i=1}^{s} w_i A_i^T \boldsymbol{\epsilon_i}.$$

This decomposition gives the optimal weights w_i and the best estimator of ζ:

$$\hat{\zeta} = \sum_{i=1}^{s} \boldsymbol{h_i}^T \boldsymbol{y_i}, \tag{5.5}$$

where $\boldsymbol{h_i} = \frac{w_i}{t}\boldsymbol{\epsilon_i}$. According to the proof of Theorem 5.1.1 indeed, an unbiased estimator of the form (5.5) is optimal if and only if every $\boldsymbol{h_i}$ is proportional to $\boldsymbol{\epsilon_i}$ and has norm $\frac{w_i}{t}$. Setting $\boldsymbol{z_i} = w_i\boldsymbol{\epsilon_i}$, one obtains t as the value of the following SOCP:

$$
\begin{aligned}
\max_{t,\boldsymbol{z},\boldsymbol{w}} \quad & t && (5.6)\\
s.t. \quad & t\boldsymbol{c} = \sum_{i=1}^{s} A_i^T \boldsymbol{z_i}, \\
& \forall i \in [s], \quad \|\boldsymbol{z_i}\| \leq w_i, \\
& \sum_i w_i = 1, \quad \boldsymbol{w} \geq \boldsymbol{0}.
\end{aligned}
$$

In order to get an SOCP in the standard form, we write $w_i = t\boldsymbol{\mu_i}$, where $t = \frac{1}{\sum_i \mu_i}$ is an arbitrary nonnegative scalar. Then, we set $\boldsymbol{h_i} = t^{-1}\boldsymbol{z_i}$, and we obtain a problem in the form

of (5.4). Finally, the value of $(P-SOCP)$ and $(D-SOCP)$ are equal, since the Slater condition holds for this pair of programs (the dual $(D-SOCP)$ is strictly feasible and the primal $(P-SOCP)$ is feasible). A proof of the strong duality theorem for SOCP can be found e.g. in [NN94], Section 4.2. See [LVBL98] for more background on SOCP duality theory. □

Remark 5.2.1. This SOCP has a simple geometric interpretation. In the scalar case, we have seen that the $\boldsymbol{c}-$optimal design could be found at the intersection of a polyhedron and a straight line directed by $\boldsymbol{c}$ (see Figure 2.2). In the multiresponse case, the generalized Elfving set is no longer a polyhedron: instead, we compute the intersection between the straight line directed by $\boldsymbol{c}$ and the set

$$\begin{aligned}\overline{\mathcal{E}} &= \operatorname{conv}\Big(\{A_i^T\boldsymbol{\epsilon_i},\ i\in[s],\ \boldsymbol{\epsilon_i}\in\mathbb{R}^{l_i},\ \|\boldsymbol{\epsilon_i}\|_2\leq 1\}\Big),\\ &= \operatorname{conv}\Big\{\mathcal{E}_i,\ i\in[s]\Big\},\end{aligned}$$

where $\mathcal{E}_i$ is the ellipsoid with semi-axis $\sqrt{\lambda_k^{(i)}}\boldsymbol{u_k^{(i)}}$ $(k\in[m])$, where $\{\lambda_1^{(i)},\dots,\lambda_m^{(i)}\}$ are the eigenvalues of $A_i^TA_i$ and $\{\boldsymbol{u_1^{(i)}},\dots,\boldsymbol{u_m^{(i)}}\}$ are the corresponding eigenvectors. In the common case, we have $l_i<m$, such that some eigenvalues of $A_i^TA_i$ vanish and the ellipsoid $\mathcal{E}_i$ is not full-dimensional (i.e. its volume is zero). We illustrate this geometric interpretation in Figure 5.1.

We next present another proof of this result, based on the rank reduction theorem of Chapter 4.

A rank reduction argument

Proof. We have seen in Chapter 3 that the $\boldsymbol{c}-$optimal design problem can be formulated as a SDP. The dual SDP (3.16) is in fact a semidefinite *packing* problem, in which the matrix defining the objective function is $C=\boldsymbol{c}\boldsymbol{c}^T$ and has rank one. Under the generic assumption that $\boldsymbol{c^T\theta}$ is *estimable*, $\boldsymbol{c}$ is in the range of $\sum_{i=1}^s A_i^TA_i$ and the conditions of Corollary 4.1.4 are fulfilled: the SDP (3.16) reduces to the SOCP (5.3).

We have seen that strong duality holds between Problems (5.3) and (5.4). This implies that any pair of primal and dual solutions $\Big(\boldsymbol{u}^*,(\boldsymbol{\mu}^*,\boldsymbol{z_i^*})\Big)$ must satisfy the complementary slackness relation

$$\forall i\in[s],\quad \mu_i^*A_i\boldsymbol{u}^*=\boldsymbol{z_i^*}.$$

Now, the dual feasibility implies that

$$\sum_i A_i\boldsymbol{z_i^*}=\sum_i\boldsymbol{\mu_i}A_i^TA_i\boldsymbol{u}^*=\boldsymbol{c}.$$

Setting $\boldsymbol{w}=t\boldsymbol{\mu}^*$ where $t^{-1}=\sum_i\mu_i^*=\boldsymbol{c}^T\boldsymbol{u}^*$, we find that $t^{-1}M(\boldsymbol{w})\boldsymbol{u}^*=\boldsymbol{c}$, and we have the equality

$$\boldsymbol{c}^TM(\boldsymbol{w})^\dagger\boldsymbol{c}=t^{-1}\boldsymbol{c}^T\boldsymbol{u}^*=(\boldsymbol{c}^T\boldsymbol{u}^*)^2.$$

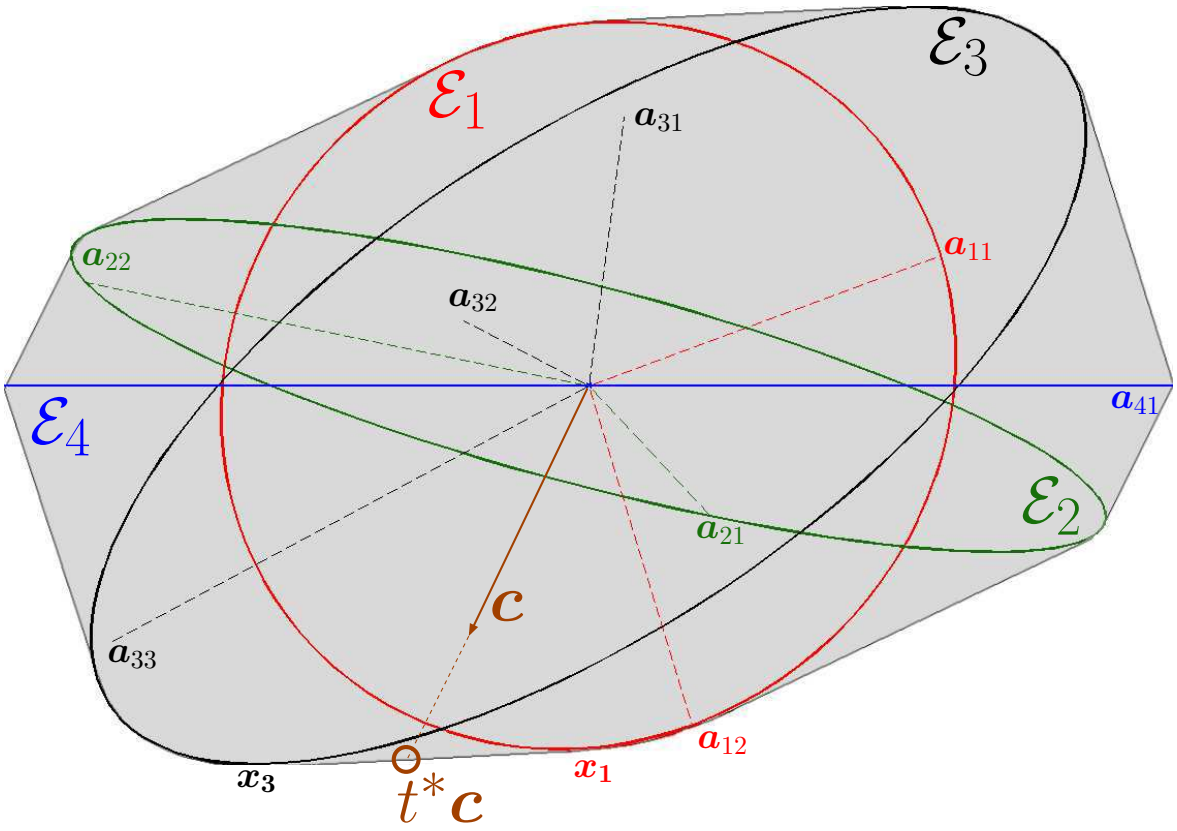


Figure 5.1: In the multiresponse case, the generalized Elfving set $\overline{\mathcal{E}}$ is the convex hull of the ellipsoids $\mathcal{E}_i$. On this picture, we have plotted the rows of the observation matrices: $\boldsymbol{a}_{ij}^T$ is the j^{th} row of A_i. In the (common) case where $l_i \leq m$, the vectors $(\boldsymbol{a}_{ij})_{j\in[l_i]}$ are on the boundary of the ellipsoid $\mathcal{E}_i$ (here, this is the case for $\mathcal{E}_1, \mathcal{E}_2$, and $\mathcal{E}_4$, but not for $\mathcal{E}_3$ since $l_3 = 3 > 2$). Note that when $l_i < m$, the ellipsoid $\mathcal{E}_i$ is not full dimensional (on the picture, we have $l_4 = 1 < 2$, so that $\mathcal{E}_4$ is a segment). The intersection of the line directed by $\boldsymbol{c}$ and the generalized Elfving set (denoted by a brown circle on the figure) indicates the weights of the $\boldsymbol{c}-$optimal design. Here, $t^*\boldsymbol{c}$ is at equal distance of the two extremal point $\boldsymbol{x}_1 \in \mathcal{E}_1$ and $\boldsymbol{x}_3 \in \mathcal{E}_3$, such that the $\boldsymbol{c}-$optimal design is $\boldsymbol{w} = [0.5, 0, 0.5, 0]^T$.

By Corollary (4.1.4), the latter expression is the optimal value of the SDP (3.16), which means that $\boldsymbol{w}$ is a $\boldsymbol{c}-$optimal design. □

Theorem 5.2.1 shows that one can compute the $\boldsymbol{c}-$optimal design on a finite regression range by solving a SOCP. This can be done very efficiently with the help of interior points codes such as SeDuMi [Stu99]. Solving the SOCP (5.3) is a much easier task than solving the SDP (3.16), because the number of variables is in the order of m (instead of m^2); because we have get rid off the positive semidefiniteness constraint of the SDP; and because the SOCP solver is able to exploit the sparse structure of the observation matrices A_i (while the partial information matrices $M_i = A_i^T A_i$ are *not very sparse* in general. Moreover, we will see in Section 5.2.3 that the SOCP approach adapts to the case of multiple linear constraints. A numerical comparison of the different algorithms that can be used to compute optimal experimental designs will be carried out in Chapter 6.

5.2.2 A-optimality

We have seen in Section 5.1.2 that any $A-$optimal design problem could be expressed as a $\boldsymbol{c}-$optimal design problem with augmented observation matrices. Thus, by Theorem 5.2.1, the $A-$optimal design problem for $K^T\boldsymbol{\theta}$ has a SOCP formulation:

Theorem 5.2.2 (Computation of the $A-$optimal design by SOCP)**.** *Let* $\left(U^*, (\boldsymbol{\mu}^*, (Z_i^*)_{i\in[s]})\right)$ *be a pair of primal and dual solutions of the second order cone programs:*

$$\begin{aligned} \max_{U\in\mathbb{R}^{m\times r}} \quad & \text{trace } K^TU \\ & \forall i \in [s], \quad \|A_iU\|_F \leq 1 \end{aligned} \tag{5.7}$$

$$\begin{aligned} \min_{\boldsymbol{\mu}\in\mathbb{R}^s,\ Z_i\in\mathbb{R}^{l_i\times r}} \quad & \sum_i \mu_i \\ & K = \sum_i A_i^T Z_i \\ & \forall i \in [s], \quad \|Z_i\|_F \leq \mu_i. \end{aligned} \tag{5.8}$$

We define

$$\boldsymbol{w} := t\boldsymbol{\mu}^*, \qquad where \qquad t = (\sum_{k=1}^{s} \mu_k^*)^{-1}.$$

Then $\boldsymbol{w}$ *is* $A-$*optimal for* $K^T\boldsymbol{\theta}$*. Moreover,* $\hat{\boldsymbol{\zeta}} = \sum_i (Z_i^*)^T \boldsymbol{y_i}$ *is the best linear estimator of* $K^T\boldsymbol{\theta}$*, and the optimal* $A-$*criterion is*

$$\Phi_A(\boldsymbol{w}) = \sum_{i=1}^{r} \boldsymbol{c_i}^T M(\boldsymbol{w}^*)^- \boldsymbol{c_i} = t^{-2} = (\sum_i \mu_i^*)^2$$

Proof. We combine the result of Section 5.1.2 and Theorem 5.2.1. □

5.2.3 c- (and A-) optimality with multiple resource constraints

In this section, we consider the generalized version of the $\boldsymbol{c}-$optimal design problem with multiple resource constraints, that we already studied in Section 3.3.5:

$$\begin{aligned} \min \quad & \boldsymbol{c}^T M(\boldsymbol{w})^- \boldsymbol{c} \\ \text{s.t.} \quad & M(\boldsymbol{w}) = A_0^T A_0 + \sum_{i=1}^{s} w_i A_i^T A_i, \\ & R\boldsymbol{w} \leq \boldsymbol{d},\ \boldsymbol{w} \geq \boldsymbol{0}. \end{aligned} \tag{5.9}$$

Note that we have added a constant $A_0^T A_0$ in the information matrix $M(\boldsymbol{w})$. This can be useful to model a *free-of-charge* experiment, that the experimenter will conduct in any case, or to model an intrinsic relationship between the parameters, such as Kirchhoff's circuit law. The constant $A_0^T A_0$ appears in $M(\boldsymbol{w})$ when we assume that the observation matrix A_0 has been normalized, in such a way that the additional observation vector $\boldsymbol{y_0}$ has a unit variance:

$$\boldsymbol{y_0} = \boldsymbol{A_0}\boldsymbol{\theta} + \boldsymbol{\varepsilon_0}, \quad \mathbb{E}[\boldsymbol{\varepsilon_0}] = \mathbf{0}, \quad \mathbb{E}[\boldsymbol{\varepsilon_0}\boldsymbol{\varepsilon_0}^T] = I. \tag{5.10}$$

Another case where it can be useful to introduce a constant term $A_0^T A_0$ in the information matrix is when a prior distribution for the parameter is given:

$$\mathbb{E}(\boldsymbol{\theta}) = \boldsymbol{\mu}, \quad \text{and} \quad \mathrm{Var}(\boldsymbol{\theta}) = R. \tag{5.11}$$

It is known (see e.g. [Puk93]) that when the prior covariance matrix R is positive definite, the expected covariance matrix is minimized (with respect to Löwner ordering) among all unbiased affine estimators, conditionally to the prior distribution of $\boldsymbol{\theta}$ for:

$$\hat{\boldsymbol{\theta}}_{|R,\boldsymbol{\mu}} = \Big(R^{-1} + \sum_{i=1}^{s} w_i A_i^T A_i\Big)^{-1}\Big(R^{-1}\boldsymbol{\mu} + \sum_{i=1}^{s} A_i^T \boldsymbol{y_i}\Big).$$

This Bayesian estimator has a variance which does not depend on the prior expected value of $\boldsymbol{\theta}$:

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}_{|R,\mu}) = \Big(R^{-1} + A(\boldsymbol{w})^T A(\boldsymbol{w})\Big)^{-1}. \tag{5.12}$$

In fact, the above discussion shows that prior information can be equivalently handled as an additional observation equation $\boldsymbol{\mu} = \boldsymbol{\theta} + \boldsymbol{\varepsilon}$, $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$, $\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = R$, which we normalize by setting $\boldsymbol{y_0} = R^{-1/2}\boldsymbol{\mu}$, $A_0 = R^{-1/2}$, $\boldsymbol{\epsilon_0} = R^{-1/2}\boldsymbol{\varepsilon}$, so that (5.10) holds. In conclusion, prior information (5.11) can be handled by adding the constant $R^{-1} = A_0^T A_0$ in the information matrix.

The main result of this section is that Problem (5.9) can be formulated as a SOCP. As in Section 5.2.1, we shall give two proofs of this result. Each proof yields a different SOCP, formulated respectively in Theorem 5.2.3 and Theorem 5.2.4. Both SOCPs are of course equivalent. We point out that a related result was obtained by Ben-Tal and Nemirovskii [BTN92], for an application to truss topology design (see also [NN94, LVBL98]).

A Statistical argument

Theorem 5.2.3. *The following SOCP is feasible if and only if $\boldsymbol{c^T\theta}$ is estimable for a feasible design ($\exists \boldsymbol{w} \geq \mathbf{0}:\ R\boldsymbol{w} \leq \boldsymbol{d}$ and $\boldsymbol{c} \in \operatorname{Im} M(\boldsymbol{w})$):*

$$\begin{aligned} \min_{\boldsymbol{w},\ \boldsymbol{\mu},\ (\boldsymbol{h_i})_{i=0,\dots,s}} \quad & \sum_{i=0}^{s} \mu_i \\ & A_0^T \boldsymbol{h_0} + \sum_{i=1}^{s} A_i^T \boldsymbol{h_i} = \boldsymbol{c} \\ & R\boldsymbol{w} \leq \boldsymbol{d},\ \boldsymbol{w} \geq \mathbf{0} \\ & \left\| \begin{bmatrix} 2\boldsymbol{h_0} \\ 1-\mu_0 \end{bmatrix} \right\| \leq 1+\mu_0 \\ & \left\| \begin{bmatrix} 2\boldsymbol{h_i} \\ w_i-\mu_i \end{bmatrix} \right\| \leq w_i+\mu_i,\ (i=1,\dots,s). \end{aligned} \tag{5.13}$$

If moreover $(\boldsymbol{w}, \boldsymbol{\mu}, (\boldsymbol{h_i})_{i=0,\dots,s})$ is a solution of Problem (5.13)*, then $\boldsymbol{w}$ is $\boldsymbol{c}-$optimal (in the sense of the general problem* (5.9)*), the best unbiased linear estimator of $\zeta = \boldsymbol{c}^T\boldsymbol{\theta}$ is $\hat{\zeta} = \sum_i \boldsymbol{h_i}^T \boldsymbol{y_i}$, and the optimal variance is $\operatorname{var}(\hat{\zeta}) = \boldsymbol{c}^T M(\boldsymbol{w})^- \boldsymbol{c} = \sum_{i=0}^{s} \mu_i$.*

Proof. The Gauss Markov Theorem 2.2.1 allows us to rewrite the objective criterion of Problem (5.9) as:

$$\boldsymbol{c}^T M(\boldsymbol{w})^- \boldsymbol{c} = \min_{\boldsymbol{h} \in \mathbb{R}^{\sum_i l_i}} \quad \boldsymbol{h}^T \Delta(\boldsymbol{w}) \boldsymbol{h} \tag{5.14}$$

$$\text{s.t.} \quad [A_0^T, A_1^T, \dots, A_s^T]\boldsymbol{h} = \boldsymbol{c}, \tag{5.15}$$

where $\Delta(\boldsymbol{w})$ is defined as in Equation (2.5), with an additional block corresponding to the prior observation ($w_0 = 1$):

$$\Delta(\boldsymbol{w}) = \begin{pmatrix} I & & & \\ & w_1^{-1} I & & \\ & & \ddots & \\ & & & w_s^{-1} I \end{pmatrix}.$$

Decomposing $\boldsymbol{h}$ as $[\boldsymbol{h_0}^T, \boldsymbol{h_1}^T, \dots, \boldsymbol{h_s}^T]^T$, $\boldsymbol{h_i} \in \mathbb{R}^{l_i}$, the expression $\boldsymbol{h}^T \Delta(\boldsymbol{w}) \boldsymbol{h}$ can be rewritten as

$$\|\boldsymbol{h_0}\|^2 + \sum_{i=1}^{s} w_i^{-1} \|\boldsymbol{h_i}\|^2. \tag{5.16}$$

Recall that when an experiment is unobserved ($w_i = 0$), it could simply be removed from the support of the experimental design. In other words, the sum (5.16) is taken on the indices such that $w_i > 0$ only. We can now rewrite Problem (5.9) in a form that involves

the vector of coefficients $\boldsymbol{h}$ of the estimator $\hat{\zeta}$:

$$\begin{aligned} \min_{\boldsymbol{w},\ (\boldsymbol{h_i}\in\mathbb{R}^{l_i})_{i=0,\ldots,s}} \quad & \|\boldsymbol{h_0}\|^2 + \sum_{\{i:w_i>0\}} \frac{\|\boldsymbol{h_i}\|^2}{w_i} \\ \text{s. t.} \quad & \sum_{i=0}^{s} A_i^T \boldsymbol{h_i} = \boldsymbol{c}, \\ & R\boldsymbol{w} \le \boldsymbol{d},\ \boldsymbol{w} \ge \boldsymbol{0}. \end{aligned} \tag{5.17}$$

Clearly, this is equivalent to:

$$\begin{aligned} \min_{\boldsymbol{w},\ \boldsymbol{\mu},\ (\boldsymbol{h_i}\in\mathbb{R}^{l_i})_{i=0,\ldots,s}} \quad & \mu_0 + \sum_{i=1}^{s} \mu_i \\ \text{s. t.} \quad & \sum_{i=0}^{s} A_i^T \boldsymbol{h_i} = \boldsymbol{c}, \\ & R\boldsymbol{w} \le \boldsymbol{d},\ \boldsymbol{w} \ge \boldsymbol{0}, \\ & \|\boldsymbol{h_0}\|^2 \le \mu_0 \\ & \|\boldsymbol{h_i}\|^2 \le \mu_i w_i, \end{aligned} \tag{5.18}$$

since we can assume without loss of generality that $w_i = 0 \Rightarrow \|h_i\| = \mu_i = 0$. Finally, the SOCP (5.13) is obtained by reformulating the hyperbolic constraints $\|\boldsymbol{z}\|^2 \le \alpha\beta$ as

$$\left\| \begin{bmatrix} 2\boldsymbol{z} \\ \alpha - \beta \end{bmatrix} \right\| \le \alpha + \beta.$$

□

A rank reduction argument

We provide another proof of the reduction of the $\boldsymbol{c}-$optimal design problem to a SOCP, which relies on the rank reduction theorem for "combined" semidefinite packing problems 4.2.2. Interestingly, we obtain a SOCP which is equivalent to (5.13) but has a different form.

Theorem 5.2.4. *The following pair of primal and dual SOCP is feasible if and only if* $\boldsymbol{c}^T\boldsymbol{\theta}$ *is estimable for a feasible design (* $\exists \boldsymbol{w} \ge \boldsymbol{0}$: $R\boldsymbol{w} \le \boldsymbol{d}$ *and* $\boldsymbol{c} \in \operatorname{Im} M(\boldsymbol{w})$*):*

$$\begin{aligned} \max_{\boldsymbol{x},\ \boldsymbol{\lambda}} \quad & \boldsymbol{c}^T\boldsymbol{x} \\ & \left\| \begin{bmatrix} 2A_0\boldsymbol{x} \\ \boldsymbol{d}^T\boldsymbol{\lambda} \end{bmatrix} \right\|_2 \leq 2 - \boldsymbol{d}^T\boldsymbol{\lambda}, \\ \forall i \in [s], \quad & \left\| \begin{bmatrix} 2A_i\boldsymbol{x} \\ \boldsymbol{r_i}^T\boldsymbol{\lambda} - 1 \end{bmatrix} \right\|_2 \leq \boldsymbol{r_i}^T\boldsymbol{\lambda} + 1 \\ & \boldsymbol{\lambda} \geq \mathbf{0}. \end{aligned} \qquad \begin{aligned} \min_{\substack{\boldsymbol{\mu}\geq\mathbf{0}, t\geq 0, (\boldsymbol{h_i})_{i=0,\dots,s} \\ \boldsymbol{\alpha}\geq\mathbf{0}, \beta\geq 0}} \quad & \sum_{i=1}^{s} \alpha_i + t + \beta \\ & A_0^T\boldsymbol{h_0} + \sum_{i=1}^{s} A_i^T\boldsymbol{h_i} = \boldsymbol{c}, \\ & R\boldsymbol{\mu} \leq t\boldsymbol{d}, \\ \forall i \in [s], \quad & \left\| \begin{bmatrix} \boldsymbol{h_i} \\ \alpha_i - \mu_i \end{bmatrix} \right\|_2 \leq \alpha_i + \mu_i, \\ & \left\| \begin{bmatrix} \boldsymbol{h_0} \\ \beta - t \end{bmatrix} \right\|_2 \leq \beta + t. \end{aligned}$$

If moreover $(\boldsymbol{\mu}, t, (\boldsymbol{h_i})_{i\in\{0,\dots,s\}}, \boldsymbol{\alpha}, \beta)$ *is a solution of the dual problem, then the optimal design variable is* $\boldsymbol{w} = t^{-1}\boldsymbol{\mu}$*, the best unbiased linear estimator of* $\zeta = \boldsymbol{c}^T\boldsymbol{\theta}$ *is* $\hat{\zeta} = \sum_i \boldsymbol{h_i}^T\boldsymbol{y_i}$*, and the optimal variance is* $\operatorname{var}(\hat{\zeta}) = \boldsymbol{c}^T M(\boldsymbol{w})^{-}\boldsymbol{c} = (\boldsymbol{c}^T\boldsymbol{x})^2 = (\sum_{i=1}^{s}\alpha_i + t + \beta)^2$.

Proof. We assume that the optimal design problem (5.9) is feasible, i.e. there exists a vector $\hat{\boldsymbol{w}} \geq \mathbf{0}$ such that $R\hat{\boldsymbol{w}} \leq \boldsymbol{d}$ and $\boldsymbol{c}$ is in the range of $M(\hat{\boldsymbol{w}})$. Note that we can assume without loss of generality that $\hat{\boldsymbol{w}} > 0$. Otherwise, this would mean that the constraints $R\boldsymbol{w} \leq \boldsymbol{d},\ \boldsymbol{w} \geq \mathbf{0}$ force the equality $w_i = 0$ to hold for some coordinate $i \in [s]$, and in this case we could simply remove the experiment i from the set of available experiments.

We can now express Problem (5.9) as an SDP thanks to the Schur complement lemma:

$$\begin{aligned} \min_{t\in\mathbb{R},\ \boldsymbol{w}\geq\mathbf{0}} \quad & t \\ \text{s.t.} \quad & \left(\begin{array}{c|c} M(\boldsymbol{w}) & \boldsymbol{c} \\ \hline \boldsymbol{c}^T & t \end{array}\right) \succeq 0. \\ & R\boldsymbol{w} \leq \boldsymbol{d}. \end{aligned} \tag{5.19}$$

Since the optimal t is positive (we exclude the trivial case $\boldsymbol{c} = \mathbf{0}$), the latter matrix inequality may be rewritten as

$$M(\boldsymbol{w}) \succeq \frac{\boldsymbol{c}\boldsymbol{c}^T}{t},$$

by using the Schur complement lemma again. Finally, we make the change of variables $\boldsymbol{\mu} = t\boldsymbol{w}$ and Problem (5.19) is equivalent to

$$\begin{aligned} \min_{\boldsymbol{\mu}\geq\mathbf{0}, t\geq 0} \quad & t \\ \text{s.t.} \quad & tA_0^TA_0 + \sum_{i=1}^{s}\mu_i A_i^TA_i \succeq \boldsymbol{c}\boldsymbol{c}^T \\ & R\boldsymbol{\mu} \leq t\boldsymbol{d}. \end{aligned} \tag{5.20}$$

This problem belongs to the class of "combined" semidefinite packing problems studied in Section 4.2. We can see indeed that Problem (5.20) has the same form as Problem (D_{CMB})

(cf. page 70), by setting $C = \boldsymbol{cc^T}$, $\mu_{s+1} = t$, $\boldsymbol{b} = [0, \dots, 0, 1]^T \in \mathbb{R}^{s+1}$, $M_{s+1} = A_0^T A_0$, $\boldsymbol{h_0} = \mathbf{0}$, $H = [R, -\boldsymbol{d}]$, and for all $i \in 0, \dots, s+1$, $R_i = 0$ (we also need to introduce a nonnegative slack variable to handle the inequalities as equalities).

Let $\lambda := \boldsymbol{c}^T (\sum_{i=0}^s M_i)^\dagger \boldsymbol{c}^T$, so that $\lambda(\sum_{i=0}^s M_i) \succeq \boldsymbol{cc}^T$. We set $\bar{t} = \max_{i \in [s]}(\lambda / \hat{w}_i, \lambda)$ ($\bar{t}$ is well defined because $\hat{\boldsymbol{w}} > \mathbf{0}$). We define $\overline{\boldsymbol{\mu}} := \bar{t}\hat{\boldsymbol{w}}$, and we see that Problem (5.20) is feasible, because $R\overline{\boldsymbol{\mu}} \leq \bar{t}\boldsymbol{d}$, and $\bar{t}M_0 + \sum_{i=1}^s \overline{\mu}_i M_i \succeq \sum_{i=0}^s \lambda M_i \succeq \boldsymbol{cc^T}$. In addition, the corresponding primal problem is clearly feasible (for $\boldsymbol{\lambda} = \mathbf{0}$, since $\boldsymbol{b} \geq \mathbf{0}$), and thus we can use Corollary 4.2.4: the $\boldsymbol{c}-$optimal design problem with resource constraints (5.9) reduces to the SOCP (4.6). With the parameters $\boldsymbol{b}, M_{s+1}, H$ and the slacks defined as above, this corresponds exactly to the primal SOCP in Theorem 5.2.4.

By construction, the optimal design variable $\boldsymbol{w}$ is related to the dual optimal variables $\boldsymbol{\mu}$ and t by the relation $\boldsymbol{w} = t^{-1}\boldsymbol{\mu}$ (according to the previous change of variable). Moreover, the dual problem satisfies the (refined) Slater condition, because $\boldsymbol{c} \in \operatorname{Im}(\sum_i M_i) = \sum_i \operatorname{Im}(A_i^T)$, so that $\exists \overline{\boldsymbol{h_0}}, \dots, \overline{\boldsymbol{h_s}} : \sum_{i=0}^s A_i^T \overline{\boldsymbol{h_i}} = \boldsymbol{c}$, $P\overline{\boldsymbol{\mu}} \leq \bar{t}\boldsymbol{d}$ and for $\overline{\boldsymbol{\alpha}} > \mathbf{0}, \overline{\beta} > 0$ large enough, the non-affine cone constraints are satisfied with a strict inequality. Hence, strong duality holds and the values of these two problems are equal. Finally, Corollary 4.2.4 shows that the optimal value of Problem (5.9) is the square of the optimal value of these SOCPs. □

5.2.4 T-optimality for $K^T\boldsymbol{\theta}$

We show in this section that it is possible to compute a *formally* $T-$optimal design for $K^T\boldsymbol{\theta}$ with a SOCP. We recall that contrarily to the other criteria of the class Φ_p , $p < 1$, a design $\boldsymbol{w}$ that maximizes $\Phi_1(\boldsymbol{w}) = \operatorname{trace} Q_K(\boldsymbol{w})$ can fail to be feasible, i.e. $\operatorname{Im} K \nsubseteq \operatorname{Im} M(\boldsymbol{w})$ (see Section 2.3.2). A *formally* $T-$optimal design $\boldsymbol{w}$ is $T-$optimal if and only if the latter range inclusion holds.

We have seen in Section 2.4.3 that the $T-$optimal design problem for the full parameter $\boldsymbol{\theta}$ is trivial: A design is formally $T-$optimal for $\boldsymbol{\theta}$ if and only if it allocates all the weight to the experiments i such that $\|A_i\|_F$ is maximal (Theorem 2.4.12). However, when the quantity of interest is a parameter subsystem $\boldsymbol{\zeta} = K^T\boldsymbol{\theta}$, the problem becomes computationally challenging. The present reduction gives another argument for saying that second order cone programming is a natural framework for optimal experimental design problems.

Theorem 5.2.5 (T-optimality SOCP)**.** *Let* $\Big((t, U), (Z_i, \boldsymbol{w}, \boldsymbol{\gamma})\Big)$ *be a pair of primal and dual*

solutions of the second order cone programs:

$$\begin{aligned} \min_{t\in\mathbb{R},\ \boldsymbol{U}\in\mathbb{R}^{m\times r}} \quad & t \\ & K^TU = I \\ & \forall i \in [s], \quad \|A_iU\|_F^2 \le t \end{aligned} \tag{5.21}$$

$$\begin{aligned} \max_{Z_0\in\mathbb{R}^{r\times r}, Z_i\in\mathbb{R}^{l_i\times r}, \boldsymbol{w}\ge\mathbf{0}, \boldsymbol{\gamma}\ge\mathbf{0}} \quad & -(\text{trace}\ \ Z_0 + \sum_{i=1}^s \gamma_i) \\ & KZ_0 = \sum_{i=1}^s A_i^T Z_i, \quad \sum_{i=1}^s w_i = 1, \\ & \forall i \in [s], \quad \|Z_i\|_F^2 \le 4w_i\gamma_i. \end{aligned} \tag{5.22}$$

Note that these are Second order cone programs indeed (we have let the hyperbolic constraints to simplify the notation; otherwise, the matrices A_iU and Z_i need be vectorized). Then, $\boldsymbol{w}$ is formally $T-$optimal for $K^T\boldsymbol{\theta}$, and the value of the supremum in Problem (2.19) *is $t = -(\text{trace}(M) + \sum_i \gamma_i)$. If moreover $\boldsymbol{w} \in \Xi(K)$, then $\boldsymbol{w}$ is $T-$optimal.*

Proof. We use the general definition (2.9) of $Q_K(\boldsymbol{w})$, which remains valid when $\boldsymbol{w} \notin \Xi(K)$:

$$\begin{aligned} Q_K(\boldsymbol{w}) := \quad & \min_{U \preceq} \quad U^TM(\boldsymbol{w})U \\ & \text{s. t.} \quad K^TU = \boldsymbol{I}_r, \end{aligned}$$

where the minimum is taken with respect to the Löwner ordering. Since the trace of a matrix preserves the Löwner ordering, we can express the (formal) $T-$optimal design problem as:

$$\begin{aligned} & \max_{\boldsymbol{w}\ge\mathbf{0},\ \sum_i w_i=1} \ \min_{U:\ K^TU=\boldsymbol{I}_r} \quad \text{trace}\, U^TM(\boldsymbol{w})U \\ = & \max_{\boldsymbol{w}\ge\mathbf{0},\ \sum_i w_i=1} \ \min_{U:\ K^TU=\boldsymbol{I}_r} \quad \sum_{i=1}^s w_i\|A_iU\|_F^2 \\ = & \min_{U:\ K^TU=\boldsymbol{I}_r} \ \left(\max_{i\in[s]} \ \|A_iU\|_F^2\right). \end{aligned}$$

The exchange of the max and the min above is a consequence of Sion's minimax theorem ($(\boldsymbol{w}, U) \mapsto \sum_{i=1}^s w_i\|A_iU\|_F^2$ is continuous, concave in $\boldsymbol{w}$ and convex in U). We next introduce a variable t which must be larger than all the quantities $\|A_iU\|_F^2$, and we have shown that the (formal) $T-$optimal design problem for $K^T\boldsymbol{\theta}$ is equivalent to Problem (5.21). The (formal) T-optimal design $\boldsymbol{w}$ is the optimal dual variable corresponding to the hyperbolic constraints in Problem (5.21). It follows that $\boldsymbol{w}$ can be computed by solving the dual optimization problem (5.22). Finally, the value of these optimization problems is the same by Strong duality (Slater condition holds), and is equal to the optimum of the $T-$optimal problem (2.19). □

5.2.5 A low rank SDP for E-optimality

Our rank reduction approach does not yield a SOCP for the computation of $E-$optimal designs. However, note that the $E-$optimality SDP (3.6) takes exactly the form of Problem (D_{PCK}) (cf. page 67), with $b_i = 1$ for all $i \in [s]$, and $C = KK^T$. Here, the matrix C has rank r, and so Theorem 4.1.2 indicates that the $E-$optimal design SDP has a solution which is a matrix of rank at most r. This suggests the use of specialized low rank solvers for this SDP when r is small (cf. the paragraph "*Related work*", page 66), which can lead to a considerable improvement in terms of computation time.

5.3 A model robust criterion

In this section, we consider the $S-$optimality criterion presented in Section 2.3.3. We are next going to show that the $S_\beta-$optimal design of multiresponse experiments reduces to the problem of maximizing a weighted geometric mean under norm constraints. This is of great interest for the computation of $S_\beta-$optimal designs. Indeed, this optimization problem is a *geometric program*, and so it can be reformulated in a form for which a self-concordant barrier function is known, and it can be solved efficiently to the desired precision by interior point techniques (see e.g. [BV04]).

Dette extended Elfving's result to the case of $S-$optimality for single-response experiments [Det93] . We will see that our result yields a generalization of the Dette's theorem for $S-$optimality to the case of multiresponse experiments. In particular, we obtain a SOCP for $D-$optimality.

5.3.1 S-optimality

We recall that the $S_\beta-$optimal design problem for the quantities $\boldsymbol{c_1}^T\boldsymbol{\theta}, \ldots, \boldsymbol{c_r}^T\boldsymbol{\theta}$ is:

$$\begin{aligned} \min \quad & S_{\boldsymbol{\beta}}(\boldsymbol{w}) := \sum_{k=1}^{r} \beta_k \log(\boldsymbol{c_k}^T M_{(k)}(\boldsymbol{w})^{-} \boldsymbol{c_k}) \\ \text{s.t.} \quad & \forall k \in [r],\ M_{(k)}(\boldsymbol{w}) = \sum_{i=1}^{s} w_i A_{(k),i}^T A_{(k),i}^T \\ & \boldsymbol{w} \geq \mathbf{0},\ \sum_{i=1}^{s} w_i \leq 1. \end{aligned} \tag{5.23}$$

The next theorem gives a *geometric programming* (GP) formulation of the $S-$optimal design problem.

Theorem 5.3.1. *Let $(\boldsymbol{t}, (\boldsymbol{v_{ik}}), \boldsymbol{w})$ be a solution of the following optimization problem. Then, $\boldsymbol{w}$ also minimizes the $S_\beta-$criterion. Moreover, the value of this program coincides*

with the value of its dual, which we give below.

$$\begin{aligned}
\min_{\boldsymbol{w}\geq\mathbf{0},\sum_i w_i=1} S_\beta(\boldsymbol{w}) = \quad & 2 \min_{\boldsymbol{t},(\boldsymbol{v_{ik}}),\boldsymbol{w}} \sum_{k=1}^{r} -\beta_k \log(t_k) \\
& t_k \boldsymbol{c_k} = \sum_{i=1}^{s} A_{(k),i}^T \boldsymbol{v_{ik}}, \qquad \forall k \in [r], \qquad (P_\beta) \\
& \left\| \begin{array}{c} \sqrt{\beta_1}\boldsymbol{v_{i1}} \\ \vdots \\ \sqrt{\beta_r}\boldsymbol{v_{ir}} \end{array} \right\| \leq w_i \qquad \forall i \in [s], \\
& \sum_{i=1}^{s} w_i \leq 1. \\
= \quad & 2 \max_{\boldsymbol{h_1},\ldots,\boldsymbol{h_r}} \sum_{k=1}^{r} \beta_k \log \frac{\boldsymbol{c_k}^T \boldsymbol{h_k}}{\beta_k} \qquad (D_\beta) \\
& \left\| \begin{array}{c} A_{(1),i}\boldsymbol{h_1}/\sqrt{\beta_1} \\ \vdots \\ A_{(r),i}\boldsymbol{h_r}/\sqrt{\beta_r} \end{array} \right\| \leq 1 \qquad \forall i \in [s].
\end{aligned}$$

The variables of the primal optimization problem are $\boldsymbol{w} \in \mathbb{R}^m$ (the design), $\boldsymbol{t} \in \mathbb{R}^r$ and the vectors $\boldsymbol{v_{ik}} \in \mathbb{R}^{l_k}$, for $i \in [s]$ and $k \in [r]$. The variables of the dual problem are the vectors $\boldsymbol{h_1}, \ldots, \boldsymbol{h_r} \in \mathbb{R}^m$.

The proof of this theorem relies on a series of reformulations of Problem (5.23) thanks to Lagrange duality techniques and Theorem 4.1.2. We will prove this result in Section 5.3.3. Then, we will show that the optimality conditions of our convex optimization problem can be interpreted as geometrical conditions, which yields a generalization of the theorem of Dette [Det93] for $S-$optimality to the case of multiresponse experiments. This geometrical characterization relies on the following generalization of the Elfving set:

$$\mathcal{D}_\beta = \operatorname{conv}\left(\left\{ \begin{pmatrix} \boldsymbol{\epsilon_1}^T A_{(1),\boldsymbol{x}} \\ \vdots \\ \boldsymbol{\epsilon_r}^T A_{(r),\boldsymbol{x}} \end{pmatrix},\ \boldsymbol{x} \in \mathcal{X},\ \boldsymbol{\epsilon_k} \in \mathbb{R}^l,\ \sum_{k=1}^{r} \beta_k \|\boldsymbol{\epsilon_k}\|^2 \leq 1 \right\} \right) \subset \mathbb{R}^{r\times m}. \quad (5.24)$$

Theorem 5.3.2 (Geometrical characterization of multiresponse $S_\beta-$optimality)**.** *The design $\boldsymbol{w}$ is $S_\beta-$optimal (and solution of Program (P_β)) if and only if there exists a vector $\boldsymbol{t} \in \mathbb{R}^r$ and vectors $\boldsymbol{\epsilon_{ik}} \in \mathbb{R}^{l_i}$ ($i \in [s], k \in [r]$), such that*

(i) $\forall i \in [s], \quad \sum_{k=1}^{r} \beta_k \|\boldsymbol{\epsilon_{ik}}\|^2 \leq 1$

(ii) $\operatorname{Diag}(\boldsymbol{t})C = \begin{pmatrix} t_1 \boldsymbol{c_1}^T \\ \vdots \\ t_r \boldsymbol{c_r}^T \end{pmatrix} = \sum_{i=1}^{s} w_i \begin{pmatrix} \boldsymbol{\epsilon_{i1}}^T A_{(1),i} \\ \vdots \\ \boldsymbol{\epsilon_{ir}}^T A_{(r),i} \end{pmatrix}$

(iii) $\mathrm{Diag}(\boldsymbol{t})C$ *lies on the boundary of* $\mathcal{D}_\beta$*, with a supporting hyperplane whose* normal direction *is given by the matrix* $H = [\boldsymbol{h_1}, \ldots, \boldsymbol{h_r}]^T$*, with* $\boldsymbol{h_k} \in \mathbb{R}^m$ *. In other words,*

$$D \in \mathcal{D}_\beta \Longrightarrow \langle H, D \rangle \leq 1$$

(iv) H *satisfies the equalities*

$$t_k \boldsymbol{h_k}^T \boldsymbol{c_k} = \beta_k, \qquad \forall k \in [r].$$

In this case, the optimal variables of Problems (D_β) *and* (P_β) *are* $\boldsymbol{t}$*,* $\boldsymbol{v_{ik}} := w_i \boldsymbol{\epsilon_{ik}}$ $(\forall i \in [s],\ \forall k \in [r])$*, and* $(\boldsymbol{h_k})_{k\in[r]}$*, so that the optimal* $S_\beta-$*criterion is* $-2\sum_{k=1}^r \beta_k \log(t_k)$.

Theorem 5.3.2 is established in the next section.

Remark 5.3.1. As in the case of single response experiments [Det93], the geometrical characterization remains true when the regression range $\mathcal{X}$ is infinite. It can also be shown with semi-infinite programming techniques that the following convex semi-infinite program is valid for the general $S_\beta-$optimal design Problem:

$$\min_{\substack{w_i \geq 0, \sum_{i=1}^s w_i = 1, \\ \boldsymbol{x} \in \mathcal{X}}} S_{\boldsymbol{\beta}}(\xi) = \quad 2 \max_{\boldsymbol{h_1}, \ldots, \boldsymbol{h_r}} \sum_{k=1}^r \beta_k \log \frac{\boldsymbol{c_k}^T \boldsymbol{h_k}}{\beta_k}$$

$$\forall \boldsymbol{x} \in \mathcal{X}, \quad \left\| \begin{array}{c} A_{(1),\boldsymbol{x}} \boldsymbol{h_1} / \sqrt{\beta_1} \\ \vdots \\ A_{(r),\boldsymbol{x}} \boldsymbol{h_r} / \sqrt{\beta_r} \end{array} \right\| \leq 1.$$

5.3.2 D-optimality

Dette showed in [Det93] that $D-$optimality for the full parameter $\boldsymbol{\theta}$ is a particular case of $S-$optimality. As a consequence, we can formulate the $D-$optimal design problem as a convex optimization problem in the form of (P_β). To see this, Dette considered the virtual nested models, where the parameter of interest in the k^{th} model is θ_k, and the observations only depend on the first k parameters: $A_{(k),i}$ is the matrix A_i restricted to its first k columns, so that $M_{(k)}(\boldsymbol{w})$ is the upper left $k \times k$ submatrix of $M(\boldsymbol{w})$, and $\boldsymbol{c_k} = [0, ..., 0, 1]$ is a vector of length k. Using the relation

$$\boldsymbol{c_k}^T M_{(k)}(\boldsymbol{w})^- \boldsymbol{c_k} = \left(M_{(k)}(\boldsymbol{w})^{-1} \right)_{kk} = \frac{\det M_{(k-1)}(\boldsymbol{w})}{\det M_{(k)}(\boldsymbol{w})},$$

it can be seen that

$$S_{[1/m,\ldots,1/m]}(\boldsymbol{w}) = -\frac{1}{m} \log\det M(\boldsymbol{w}),$$

which is exactly the $D-$optimality criterion.
Theorem 5.3.1 can now be used to formulate the $D-$optimal design problem as:

$$\begin{aligned}\max_{\boldsymbol{w}\geq\mathbf{0}} \frac{1}{m}\log\det M(\boldsymbol{w}) = \quad & 2\max_{\boldsymbol{t},\boldsymbol{v_{ik}},\boldsymbol{w}} \log\left(\left(\prod t_i\right)^{1/m}\right) \\ & t_k\boldsymbol{c_k} = \sum_i A_{(k),i}^T \boldsymbol{v_{ik}}, \qquad \forall k\in[m], \\ & \left\| \begin{array}{c} \boldsymbol{v_{i1}} \\ \vdots \\ \boldsymbol{v_{im}} \end{array} \right\| \leq \sqrt{m}\; w_i \qquad \forall i\in[s], \\ & \sum_{i=1}^s w_i \leq 1.\end{aligned} \tag{5.25}$$

5.3.3 Proof of Theorems 5.3.1 and 5.3.2

We start with the following lemma, where we show that the $S_\beta-$optimal design problem can be formulated as a convex optimization problem with SDP constraints:

Lemma 5.3.3. *The optimal variable $\boldsymbol{w}^*$ of the following convex optimization problem also minimizes the $S_\beta-$criterion. The value of this program coincides with the value of its dual, which we give below:*

$$\begin{aligned}\min_{\boldsymbol{w}\geq\mathbf{0},\sum_i w_i=1} S_\beta(\boldsymbol{w}) = \min_{\boldsymbol{\tau}\in\mathbb{R}^r,\ \boldsymbol{w}\geq\mathbf{0}} & -\sum_{k=1}^r \beta_k\log\tau_k \qquad (P_\beta - SDP) \\ & M_{(k)}(\boldsymbol{w}) \succeq \tau_k \boldsymbol{c_k}\boldsymbol{c_k}^T, \qquad \forall k\in[r], \\ & \sum_{i=1}^s w_i = 1.\end{aligned}$$

$$\begin{aligned}= \max_{Z_1,\ldots,Z_r\succeq 0} & \sum_{k=1}^r \beta_k \log\frac{\boldsymbol{c_k}^T Z_k \boldsymbol{c_k}}{\beta_k} \qquad (D_\beta - SDP) \\ & \sum_{k=1}^r \operatorname{trace}(A_{(k),i}\, Z_k\, A_{(k),i}^T) \leq 1, \qquad \forall i\in[s].\end{aligned}$$

Proof. As in the derivation of the SDP for $A-$optimality (cf. page 59), we reexpress the variance of the k^{th} quantity of interest $\boldsymbol{c_k}^T M_{(k)}(\boldsymbol{w})^- \boldsymbol{c_k}$ with the help of a generalized Schur complement (for an arbitrary design $\boldsymbol{w}$):

$$\left(\boldsymbol{c_k}^T M_{(k)}(\boldsymbol{w})^- \boldsymbol{c_k}\right)^{-1} = \begin{array}{c}\max \quad \tau_k \\ \left(\begin{array}{c|c} M_{(k)}(\boldsymbol{w}) & \boldsymbol{c_k} \\ \hline \boldsymbol{c_k}^T & 1/\tau_k \end{array}\right) \succeq 0.\end{array} \qquad = \begin{array}{c}\max \quad \tau_k \\ M_{(k)}(\boldsymbol{w}) \succeq \tau_k \boldsymbol{c_k}\boldsymbol{c_k}^T.\end{array}$$

Since the optimal τ_k is positive, the latter expression is well defined. Now, by monotonicity of the log function, we can write:

$$\begin{aligned}\min_{\boldsymbol{w}\geq\mathbf{0},\sum_i w_i=1} S_\beta(\boldsymbol{w}) &= -\max_{\boldsymbol{w}\geq\mathbf{0},\sum_i w_i=1} \quad \sum_{k=1}^{r}\beta_k\log\Big(\boldsymbol{c_k}^T M_{(k)}(\boldsymbol{w})^{-}\boldsymbol{c_k}\Big)^{-1}\\ &= -\max_{\boldsymbol{w}\geq\mathbf{0},\sum_i w_i=1,\ \boldsymbol{\tau}\in\mathbb{R}^r} \quad \sum_{k=1}^{r}\beta_k\log\tau_k\\ &\qquad\qquad M_{(k)}(\boldsymbol{w})\succeq\tau_k\boldsymbol{c_k}\boldsymbol{c_k}^T, \qquad \forall k\in[r],\end{aligned}$$

which is exactly Problem $(P_\beta - SDP)$. It is clear that Problem $(D_\beta - SDP)$ is convex and strictly feasible, so that the Slater condition is fulfilled, and strong duality holds. It remains to show that Problem $(D_\beta - SDP)$ is indeed the dual of $(P_\beta - SDP)$. To this end, let us form the Lagrangian of Problem $(P_\beta - SDP)$:

$$\mathcal{L}\Big((\boldsymbol{\tau},\boldsymbol{w}),(Z,\lambda)\Big) = -\sum_{k=1}^{r}\beta_k\log\tau_k + \sum_{k=1}^{r}\langle Z_k, \tau_k\boldsymbol{c_k}\boldsymbol{c_k}^T - M_{(k)}(\boldsymbol{w})\rangle + \lambda(\sum_{i=1}^{s}w_i - 1).$$

The Lagrange dual function is given by

$$\begin{aligned}g(Z,\lambda) &:= \min_{\boldsymbol{\tau}>\mathbf{0},\ \boldsymbol{w}\geq\mathbf{0}} \ \mathcal{L}\Big((\boldsymbol{\tau},\boldsymbol{w}),(Z,\lambda)\Big)\\ &= -\lambda + \sum_k \min_{\tau_k>0}(\tau_k\boldsymbol{c_k}^T Z_k\boldsymbol{c_k} - \beta_k\log\tau_k) + \sum_i \min_{\boldsymbol{w}\geq\mathbf{0}}\ w_i(\lambda - \sum_k\langle A_{(k),i}^T\ A_{(k),i}, Z_k\rangle).\\ &= \begin{cases} -\lambda + \sum_k\beta_k(1-\log\frac{\beta_k}{\boldsymbol{c_k}^T Z_k\boldsymbol{c_k}}) & \text{if}\begin{cases}\forall i\in[s],\ \sum_{k=1}^{r}\langle A_{(k),i}^T\ A_{(k),i}, Z_k\rangle\leq\lambda\\ \forall k\in[r],\ \boldsymbol{c_k}^T Z_k\boldsymbol{c_k}>0\end{cases}\\ -\infty & \text{otherwise.}\end{cases}\end{aligned}$$

Note that in the above expression, the minimum over τ_k is attained for $\tau_k = \frac{\beta_k}{\boldsymbol{c_k}^T Z_k\boldsymbol{c_k}}$, and this equation must be satisfied by the optimal variables τ_k^* and Z_k^*. Since we observed that strong duality holds, the value of the dual optimization problem must be equal to the value of the primal, and so the optimal variables (denoted with stars in superscript) satisfy:

$$-\sum_{k=1}^{r}\beta_k\log\tau_k^* = -\lambda^* + \sum_k\beta_k(1-\log\frac{\beta_k}{\boldsymbol{c_k}^T Z_k^*\boldsymbol{c_k}}) \qquad \Longrightarrow \lambda^* = \sum_{k=1}^{r}\beta_k = 1.$$

We can now make the dual problem explicit:

$$\begin{aligned}\max_{Z,\lambda} g(Z,\lambda) = \max_{Z_1,\dots,Z_r\succeq 0} g(Z,1) = \max_{Z_1,\dots Z_s\succeq 0} \quad & \sum_{k=1}^{r}\beta_k\log\frac{\boldsymbol{c_k}^T Z_k\boldsymbol{c_k}}{\beta_k}\\ & \sum_{k=1}^{r}\mathrm{trace}(A_{(k),i}\ Z_k\ A_{(k),i}^T)\leq 1, \qquad \forall i\in[s].\end{aligned}$$

This completes the proof of the lemma. □

Now, we show that there is a solution of Problem $(D_\beta - SDP)$ for which every Z_k has

rank one, thanks to the theoretical result of Chapter 4

Proof of Theorem 5.3.1. We first write the program $(D_\beta - SDP)$ in the form of a separable optimization problem, by introducing some vectors $\boldsymbol{\alpha_i}$ $(i \in [s])$ of size r, satisfying $\sum_{k=1}^{r} \alpha_{ik} \leq 1$:

$$\min_{\boldsymbol{w}\geq \mathbf{0}, \sum_i w_i = 1} S_\beta(\boldsymbol{w}) = \max_{\boldsymbol{\alpha_1}, \ldots \boldsymbol{\alpha_s} \in \mathbb{R}^r} \left(\sum_{k=1}^{r} f_k(\alpha_{1k}, \ldots, \alpha_{sk}) \right)$$
$$\forall i \in [s],\ \sum_{k=1}^{r} \alpha_{ik} \leq 1,$$

where we have set

$$\forall k \in [r], \qquad f_k(\boldsymbol{y}) = \max_{Z_k \succeq 0} \quad \beta_k \log \frac{\boldsymbol{c_k}^T Z_k \boldsymbol{c_k}}{\beta_k}$$
$$\operatorname{trace}(A_{(k),i}\ Z_k\ A_{(k),i}^T) \leq y_i, \quad \forall i \in [s].$$

By use of Theorem (4.1.2) (and monotonicity of the log function), the minimization problem over Z_k in $f_k(\cdot)$ has a rank-one solution ($Z_k = \boldsymbol{h_k}\boldsymbol{h_k}^T$), and we obtain:

$$f_k(\alpha_{1k}, \ldots, \alpha_{sk}) = \max_{\boldsymbol{h_k} \in \mathbb{R}^m} \quad \beta_k \log \frac{(\boldsymbol{c_k}^T \boldsymbol{h_k})^2}{\beta_k}$$
$$\|A_{(k),i}\ \boldsymbol{h_k}\| \leq \sqrt{\alpha_{ik}}, \quad \forall i \in [s].$$

Now, we use the associativity of the maximum to reformulate the $S_\beta-$optimum design problem:

$$\min_{\boldsymbol{w}\geq \mathbf{0}, \sum_i w_i = 1} S_\beta(\boldsymbol{w}) = \quad \max_{\boldsymbol{h_1}, \ldots, \boldsymbol{h_s}} \sum_{k=1}^{r} \beta_k \log \frac{(\boldsymbol{c_k}^T \boldsymbol{h_k})^2}{\beta_k}$$
$$\sum_{k=1}^{r} \|A_{(k),i}\ \boldsymbol{h_k}\|^2 \leq 1, \quad \forall i \in [s].$$

Finally, we make the change of variable $\boldsymbol{h_k}' = \boldsymbol{h_k}\sqrt{\beta_k}$ in order to obtain the desired optimization problem, that is (D_β). It remains to show that Problem (P_β) is the dual of (D_β). The convex problem (P_β) is strictly feasible, so that Slater condition is fulfilled, and strong duality holds.
We will now dualize Problem (P_β). This part of the proof is very similar to the dualization of Problem $(D_\beta - SDP)$ of the previous lemma. We include it here, though, for the reader's convenience. In the sequel, we denote by $\boldsymbol{v_i}$ the concatenation of the vectors $\boldsymbol{v_{ik}}$: $\boldsymbol{v_i} = [\boldsymbol{v_{i1}}^T, \ldots, \boldsymbol{v_{ir}}^T]^T \in \mathbb{R}^{rl}$, and by $\tilde{\boldsymbol{\beta}}$ the vector containing β_k entries arranged in blocks of length l: $\tilde{\boldsymbol{\beta}} = [\beta_1, \ldots, \beta_1, \ldots, \ldots, \beta_r, \ldots, \beta_r]^T \in \mathbb{R}^{rl}$. We also use the symbol $\odot$

to denote the Hadamard product (elementwise product). With this notation, we can write:

$$\begin{pmatrix} \sqrt{\beta_1}\boldsymbol{v_{i1}} \\ \vdots \\ \sqrt{\beta_r}\boldsymbol{v_{ir}} \end{pmatrix} = \tilde{\boldsymbol{\beta}}^{1/2} \odot \boldsymbol{v_i}.$$

We denote by V the family of vectors $(\boldsymbol{v_{ik}})_{i\in[s],\ k\in[r]}$ and by H the family of vectors $(\boldsymbol{h_k})_{i\in[s]}$. Now, let us form the Lagrangian

$$\begin{aligned}\mathcal{L}\Big((\boldsymbol{t},V,\boldsymbol{w}),(H,\boldsymbol{\mu},\lambda)\Big) = \sum_{k=1}^{r} -\beta_k \log t_k + \sum_{k=1}^{r} \boldsymbol{h_k}^T(t_k\boldsymbol{c_k} - \sum_{i=1}^{s} A^T_{(k),i}\boldsymbol{v_{ik}}) \\ + \sum_{i=1}^{s} \mu_i(\|\tilde{\boldsymbol{\beta}}^{1/2} \odot \boldsymbol{v_i}\| - w_i) + \lambda(\sum_{i=1}^{s} w_i - 1)\end{aligned} \tag{5.26}$$

The Lagrange dual function is given by

$$\begin{aligned} g(H,\boldsymbol{\mu},\lambda) &:= \min_{\boldsymbol{t},V,\ w} \mathcal{L}\Big((\boldsymbol{t},V,\boldsymbol{w}),(H,\boldsymbol{\mu},\lambda)\Big) \\ &= -\lambda + \sum_{k=1}^{r} \min_{t_k}(t_k \boldsymbol{h_k}^T\boldsymbol{c_k} - \beta_k \log t_k) + \sum_{i=1}^{s} \min_{w_i} w_i(\lambda - \mu_i) \\ &\quad + \sum_{i=1}^{s} \min_{\boldsymbol{v_i}}(\mu_i \|\tilde{\boldsymbol{\beta}}^{1/2} \odot \boldsymbol{v_i}\| - \boldsymbol{z_i}^T\boldsymbol{v_i}), \end{aligned}$$

where we have defined the vectors $\boldsymbol{z_i}^T := [\boldsymbol{h_1}^T A^T_{(1),i}, ..., \boldsymbol{h_r}^T A^T_{(r),i}] \in \mathbb{R}^{rl}$. In the latter equation, the minimum over t_k is finite if and only if $\boldsymbol{c_k}^T\boldsymbol{h_k} > 0$, and is attained for $t_k = \frac{\beta_k}{\boldsymbol{c_k}^T\boldsymbol{h_k}}$; the expression in w_i is bounded from below (by 0) if and only if $\mu_i = \lambda$. The reader can also verify that the minimization with respect to $\boldsymbol{v_i}$ is unbounded whenever $\|\tilde{\boldsymbol{\beta}}^{-1/2} \odot \boldsymbol{z_i}\| > \mu_i$, and takes the value 0 otherwise. The Cauchy Schwarz inequality between the vectors $\tilde{\boldsymbol{\beta}}^{-1/2} \odot \boldsymbol{z_i}$ and $\tilde{\boldsymbol{\beta}}^{1/2} \odot \boldsymbol{v_i}$ shows indeed that the minimum is attained for a vector such that $\boldsymbol{v_i}$ is proportional to $\tilde{\boldsymbol{\beta}}^{-1} \odot \boldsymbol{z_i}$ if $\|\tilde{\boldsymbol{\beta}}^{-1/2} \odot \boldsymbol{z_i}\| = \mu_i$, and for $\boldsymbol{v_i} = \boldsymbol{0}$ if $\|\tilde{\boldsymbol{\beta}}^{-1/2} \odot \boldsymbol{z_i}\| < \mu_i$. To summarize,

$$g(H,\boldsymbol{\mu},\lambda) = \begin{cases} -\lambda + \sum_{k=1}^{r} \beta_k(1 - \log \frac{\beta_k}{\boldsymbol{c_k}^T\boldsymbol{h_k}}) & \text{if } \forall i \in [s],\ \mu_i = \lambda \text{ and } \|\tilde{\boldsymbol{\beta}}^{-1/2} \odot \boldsymbol{z_i}\| \leq \mu_i; \\ -\infty & \text{otherwise.} \end{cases}$$

Now, since the primal and the dual share the same optimal value (we observed that strong duality holds), it follows that the optimal variables (denoted with stars in superscript) satisfy

$$g(H^*,\boldsymbol{\mu^*},\lambda^*) = -\lambda^* + \sum_{k=1}^{r} \beta_k(1 - \log \frac{\beta_k}{\boldsymbol{c_k}^T\boldsymbol{h_k}^*}) = \sum_{k=1}^{r} -\beta_k \log t_k^*.$$

Combining this equality with the stationarity equations $t_k^* = \frac{\beta_k}{\boldsymbol{c_k}^T \boldsymbol{h_k}^*}$ and $\mu_i^* = \lambda^*$, we obtain:

$$\lambda^* = \mu_i^* = \sum_{k=1}^{r} \beta_k = 1, \qquad \forall i \in [s].$$

We can now make the dual of (P_β) explicit:

$$\begin{aligned} \min_{\boldsymbol{w}\geq \mathbf{0}, \sum w_i = 1} S_{\boldsymbol{\beta}}(\boldsymbol{w}) = \quad & 2 \max_{H} \sum_{k=1}^{r} \beta_k \log \frac{\boldsymbol{c_k}^T \boldsymbol{h_k}}{\beta_k} \\ & \boldsymbol{z_i} = \begin{pmatrix} A_{(1),i} \boldsymbol{h_1} \\ \vdots \\ A_{(r),i} \boldsymbol{h_r} \end{pmatrix}, \quad \forall i \in [s], \\ & \|\tilde{\boldsymbol{\beta}}^{-1/2} \odot \boldsymbol{z_i}\| \leq 1, \qquad \forall i \in [s]. \end{aligned}$$

This program is the same as (D_β), and it completes the proof of Theorem 5.3.1. □

Now, we can write that a design is optimal if and only if Karush Kuhn Tucker (KKT) optimality conditions hold for problem (P_β). In fact, we show in Theorem 5.3.2 that these KKT conditions are equivalent to a geometrical characterization of $S_\beta-$optimality, which generalizes the theorem of Dette [Det93] to the case of multiresponse experiments.

Proof of Theorem 5.3.2. Since strong duality holds between Problems (P_β) and (D_β), the Karush Kuhn Tucker (KKT) conditions characterize the optimal variables. We sum up the KKT conditions here, which stem from the dualization step of the proof of Theorem 5.3.1:

$$\text{(Feasibility)} \qquad t_k \boldsymbol{c_k} = \sum_{i=1}^{s} A_{(k),i}^T \, \boldsymbol{v_{ik}} \tag{5.27}$$

$$\sum_{i=1}^{s} w_i = 1 \tag{5.28}$$

$$\text{(Comp. Slackness)} \qquad \mu_i(\|\tilde{\boldsymbol{\beta}}^{1/2} \odot \boldsymbol{v_i}\| - w_i) = \mathbf{0} \overset{(\text{since } \mu_i = 1)}{\Longrightarrow} w_i = \|\tilde{\boldsymbol{\beta}}^{1/2} \odot \boldsymbol{v_i}\| \tag{5.29}$$

$$\text{(Stationarity)} \qquad \beta_k = t_k \boldsymbol{h_k}^T \boldsymbol{c_k} \tag{5.30}$$

$$\begin{cases} \|\tilde{\boldsymbol{\beta}}^{-1/2} \odot \boldsymbol{z_i}\| \leq 1 \text{ and } \boldsymbol{v_i} = \mathbf{0} & \text{if } w_i = 0 \\ \|\tilde{\boldsymbol{\beta}}^{-1/2} \odot \boldsymbol{z_i}\| = 1 \text{ and } \boldsymbol{v_i} = w_i \, \tilde{\boldsymbol{\beta}}^{-1} \odot \boldsymbol{z_i} & \text{otherwise.} \end{cases} \tag{5.31}$$

In the above equations, the vector $\boldsymbol{z_i}$ is used to denote the vector $[\boldsymbol{h_1}^T A_{(1),i}^T, ..., \boldsymbol{h_r}^T A_{(r),i}^T]^T \in \mathbb{R}^{rl}$. Now, let $(\boldsymbol{t}, V, \boldsymbol{w})$ and $H = [\boldsymbol{h_1}, ..., \boldsymbol{h_r}]^T$ be a pair of primal and dual solutions of Problem (P_β)–(D_β): they satisfy KKT equations(5.27)-(5.31). We set $\boldsymbol{\epsilon_i} = \frac{1}{w_i} \boldsymbol{v_i}$ whenever $w_i \neq 0$ and $\boldsymbol{\epsilon_i} = \mathbf{0} \in \mathbb{R}^{rl}$ otherwise, so that (5.29) implies

$$\forall i \in [s], \quad \sum_{k=1}^{r} \beta_k \|\boldsymbol{\epsilon_{ik}}\|^2 = w_i \leq 1$$

and (5.27) implies

$$\forall k \in [r], \quad t_k \boldsymbol{c_k} = \sum_{i=1}^{s} A_{(k),i}^T \boldsymbol{v_{ik}} = \sum_{i=1}^{s} w_i A_{(k),i}^T \boldsymbol{\epsilon_{ik}}.$$

These relations are nothing but conditions (i) and (ii) of Theorem (5.3.2). Clearly, the stationarity equation (5.30) is the same as condition (iv) of Theorem (5.3.2). It remains to show that (iii) holds. Let D be an arbitrary matrix from the generalized Elfving set $\mathcal{D}_\beta$ (cf. Equation (5.24)). When the regression region is $\mathcal{X} = [s]$, there exists a vector $\boldsymbol{\alpha}$ in the unit simplex of $\mathbb{R}^s$ as well as vectors $(\boldsymbol{\delta_i} := [\boldsymbol{\delta_{i1}}^T, \ldots, \boldsymbol{\delta_{ir}}^T]^T \in \mathbb{R}^{rl})_{i\in[s]}$ satisfying $\|\tilde{\boldsymbol{\beta}}^{1/2} \odot \boldsymbol{\delta_i}\| \leq 1$ such that

$$D = \sum_{i=1}^{s} \alpha_i \begin{pmatrix} \boldsymbol{\delta_{i1}}^T A_{(1),i} \\ \vdots \\ \boldsymbol{\delta_{ir}}^T A_{(r),i} \end{pmatrix}.$$

We now prove that $H = [\boldsymbol{h_1}, ..., \boldsymbol{h_r}]^T$ is the *direction* of the supporting hyperplane of $\mathcal{D}_\beta$:

$$\begin{aligned} \langle D, H \rangle &= \sum_{i,k} \alpha_i \boldsymbol{\delta_{ik}}^T A_{(k),i} \boldsymbol{h_k} \\ &= \sum_{i=1}^{s} \alpha_i \boldsymbol{\delta_i}^T \boldsymbol{z_i} \\ &= \sum_{i=1}^{s} \alpha_i (\tilde{\boldsymbol{\beta}}^{1/2} \odot \boldsymbol{\delta_i})^T (\tilde{\boldsymbol{\beta}}^{-1/2} \odot \boldsymbol{z_i}) \\ &\leq \sum_{i=1}^{s} \alpha_i \leq 1, \end{aligned}$$

where the inequality is Cauchy-Schwarz, and we have used the stationarity condition (5.31). Finally, (iii) holds since $\mathrm{Diag}(\boldsymbol{t})C$ lies on the boundary of $\mathcal{D}_\beta$:

$$\langle \mathrm{Diag}(\boldsymbol{t})C, H \rangle = \sum_{k=1}^{r} t_k \boldsymbol{c_k}^T \boldsymbol{h_k} = \sum_k \beta_k = 1.$$

Conversely, assume that conditions $(i) - (iv)$ hold. We set $\boldsymbol{v_i} = w_i \boldsymbol{\epsilon_i}$, and we show that $(\boldsymbol{t}, V, \boldsymbol{w})$ and H satisfy the KKT equations(5.27)-(5.31). As in the direct part of this proof, it is straightforward to show that the stationarity equation (5.30) holds, as well as the feasibility condition (5.27).
Let us now define the vector $\boldsymbol{z_i}$ as in (D_β):

$$\boldsymbol{z_i} = \begin{pmatrix} A_{(1),i} \boldsymbol{h_1} \\ \vdots \\ A_{(r),i} \boldsymbol{h_r} \end{pmatrix}.$$

Condition (iii) states that for all vector $\boldsymbol{\alpha}$ in the unit simplex of $\mathbb{R}^s$, and for all vectors

$(\boldsymbol{\delta_i} \in \mathbb{R}^{sl})_{i\in[s]}$ satisfying $\|\tilde{\boldsymbol{\beta}}^{1/2} \odot \boldsymbol{\delta_i}\| \leq 1$, we have

$$\sum_{ik} \alpha_i \boldsymbol{\delta_{ik}}^T A_{(k),i} \boldsymbol{h_k} \leq 1.$$

In particular, if $\boldsymbol{\alpha} = \boldsymbol{e_i}$ is the i^{th} unit vector of the canonical basis of $\mathbb{R}^s$, and $\boldsymbol{\delta_i} = \frac{\tilde{\boldsymbol{\beta}}^{-1} \odot \boldsymbol{z_i}}{\|\tilde{\boldsymbol{\beta}}^{-1/2} \odot \boldsymbol{z_i}\|}$, we obtain:

$$\sum_{ik} \alpha_i \boldsymbol{\delta_{ik}}^T A_{(k),i} \boldsymbol{h_k} = \boldsymbol{\delta_i}^T \boldsymbol{z_i} = \frac{1}{\|\tilde{\boldsymbol{\beta}}^{-1/2} \odot \boldsymbol{z_i}\|} (\tilde{\boldsymbol{\beta}}^{-1/2} \odot \boldsymbol{z_i})^T (\tilde{\boldsymbol{\beta}}^{-1/2} \odot \boldsymbol{z_i}) = \|\tilde{\boldsymbol{\beta}}^{-1/2} \odot \boldsymbol{z_i}\| \leq 1,$$

and we have shown the inequality of (5.31).
The fact that $\boldsymbol{v_i} = 0$ when $w_i = 0$ is clear from the way we have defined $\boldsymbol{v_i}$, and the complementarity slackness equation (5.29) also holds in this case.
It remains to show that $\boldsymbol{w}$ is feasible (5.28) and that (5.31) holds for $w_i > 0$. Note that (5.31) in turn implies the complementarity slackness equation (5.29).

To this end, we write:

$$\begin{aligned} 1 = \sum_{k=1}^{r} \beta_k = \sum_{k=1}^{r} t_k \boldsymbol{c_k}^T \boldsymbol{h_k} &= \sum_{ik} w_i \boldsymbol{\epsilon_{ik}}^T A_{(k),i} \boldsymbol{h_k} \\ &= \sum_{i=1}^{s} w_i \boldsymbol{\epsilon_i}^T \boldsymbol{z_i} \\ &= \sum_{i=1}^{s} w_i (\tilde{\boldsymbol{\beta}}^{1/2} \odot \boldsymbol{\epsilon_i})^T (\tilde{\boldsymbol{\beta}}^{-1/2} \odot \boldsymbol{z_i}) \\ &\leq \sum_{i=1}^{s} w_i \|\tilde{\boldsymbol{\beta}}^{1/2} \odot \boldsymbol{\epsilon_i}\| \ \|\tilde{\boldsymbol{\beta}}^{-1/2} \odot \boldsymbol{z_i}\|. \end{aligned}$$

The latter inequality is Cauchy-Schwarz, and it provides an upper bound which is the (weighted) mean of terms all smaller than 1. We can therefore write

$$\sum_{i=1}^{s} w_i \|\tilde{\boldsymbol{\beta}}^{1/2} \odot \boldsymbol{\epsilon_i}\| \ \|\tilde{\boldsymbol{\beta}}^{-1/2} \odot \boldsymbol{z_i}\| = 1, \tag{5.32}$$

and the Cauchy-Schwarz inequality must be an equality whenever $w_i \neq 0$, which occurs if and only if $\tilde{\boldsymbol{\beta}}^{1/2} \odot \boldsymbol{\epsilon_i}$ is proportional to $\tilde{\boldsymbol{\beta}}^{-1/2} \odot \boldsymbol{z_i}$. Finally, we must have $\sum_i w_i = 1$, so that $\boldsymbol{w}$ is feasible (5.28), and each positively weighted term in the sum (5.32) must be 1:

$$w_i \neq 0 \Longrightarrow \|\tilde{\boldsymbol{\beta}}^{1/2} \odot \boldsymbol{\epsilon_i}\| \ \|\tilde{\boldsymbol{\beta}}^{-1/2} \odot \boldsymbol{z_i}\| = 1 \Longrightarrow \begin{cases} \|\tilde{\boldsymbol{\beta}}^{1/2} \odot \boldsymbol{\epsilon_i}\| = 1 \\ \|\tilde{\boldsymbol{\beta}}^{-1/2} \odot \boldsymbol{z_i}\| = 1 \end{cases}.$$

These two norm constraints further force the coefficient of proportionality between $\tilde{\boldsymbol{\beta}}^{1/2} \odot \boldsymbol{\epsilon_i}$ and $\tilde{\boldsymbol{\beta}}^{-1/2} \odot \boldsymbol{z_i}$ to be 1, so that $\boldsymbol{\epsilon_i} = \tilde{\boldsymbol{\beta}}^{-1} \odot \boldsymbol{z_i}$, and $\boldsymbol{v_i} = w_i \tilde{\boldsymbol{\beta}}^{-1} \odot \boldsymbol{z_i}$, which completes the proof. □

Chapter 6

Numerical comparison of the algorithms

In this chapter, we compare the numerical performance of the different algorithms discussed in the previous chapters. We will see that the second order cone programs presented in this chapter are very efficient when the number r of quantities to estimate is small (in particular for $\boldsymbol{c}-$optimality).

We will compare our approach to the classic algorithms presented in Chapter 3. In particular, we concentrate on the semidefinite programming/MAXDET approach [VBW98], Wynn–Fedorov-type exchange algorithms [Wyn70, Fed72], and Titterington-type multiplicative algorithms [Tit76]. Several versions and refinements of these procedures were proposed. For the class of exchange algorithms, we will use the *IncDec* procedure of Richtarik [Ric08], which specifies step lengths for which the precision δ is achieved in $O(1/\delta)$ iterations; for the multiplicative algorithms, we will use the general class of iterations introduced by Silvey, Titterington and Torsney [STT78], which is defined by a power parameter λ (cf. Equation (3.3)) and is known to converge to an optimal design under certain conditions [Yu10a]. We will also consider a variant of the latter algorithm which uses an acceleration parameter γ, for which Dette, Pepelyshev and Zhigljavsky [DPZ08] have established a convergence result in the case of $D-$optimality, and conjectured the convergence for other criteria. We found that the values $\lambda = 0.9$ and $\gamma = 0.9$ gave the best results for $A-$optimality in our experiments, and so those values will be used throughout this chapter. For $D-$optimality, we have used the acceleration parameter $\gamma = 0.5$.

We will first consider random instances of optimal design problems, in order to evaluate to which extent each parameter affects the computation time. Then, we will consider a simple polynomial regression model, for which we shall see that our approach is well-suited when the number of support points is large. Finally we will present some results from the network application which we be detailed in Chapter 10, where the sampling rates of a monitoring tool should be optimized subject to multiple constraints.

m	SOCP (5.7) [this paper]	SDP [VBW98]	IncDec Exchange [Ric08]	Accelerated Mult. algo ($\gamma = 0.9$) [DPZ08]	Mult. algo with Exponent $\lambda = 0.9$ [Yu10a]
2	0.082	2.897	10.039	3.026	2.979
2^2	0.120	3.017	99.510	9.598	9.240
2^3	0.166	4.798	13.112	5.883	6.040
2^4	0.175	6.828	24.431	12.574	12.204
2^5	0.352	15.820	29.454	11.258	11.123
2^6	0.816	66.281	54.379	13.407	13.419
2^7	2.636	338.669	92.537	37.935	36.679
2^8	10.496	failed	202.509	96.594	99.751
2^9	44.689	failed	412.890	585.619	597.442
2^{10}	154.187	failed	498.616	551.634	539.130

Table 6.1: CPU time (s) of the different algorithms, for typical random instances of the $A-$optimal design problem with $s = 2^{10}$, $l = 1$, $r = 3$, and different values of m.

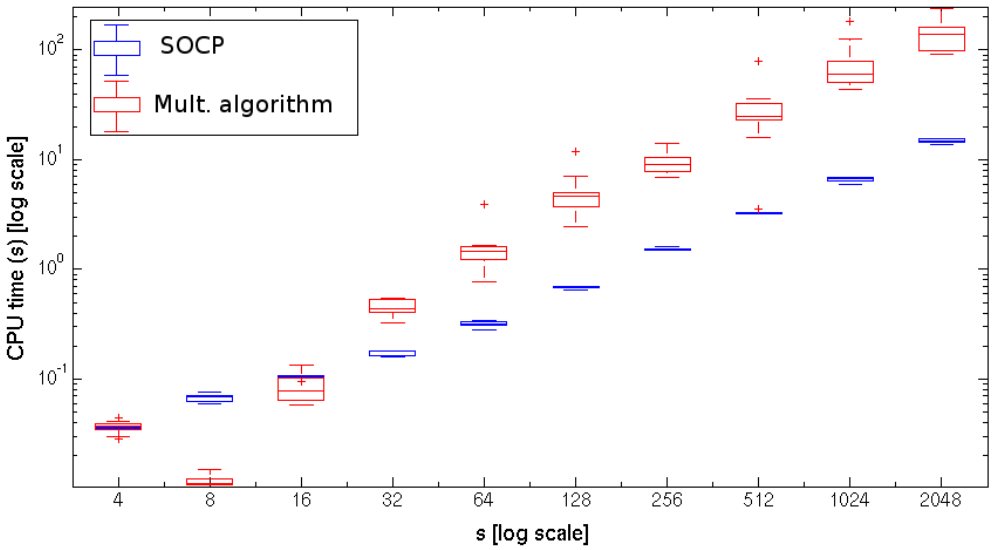


Figure 6.1: Comparison of two algorithms (SOCP vs. multiplicative algorithm with the acceleration parameter $\gamma = 0.9$ [DPZ08]) on random instances ($A-$optimality) with $m = 120$, $l = 30$, $r = 1$, and varying s. The box plots represent the distribution of the computing times for 10 random instances.

6.1 Random instances

In this section, we consider random instances of optimal experimental design problems, in which the entries of the $l \times m$ matrices $(A_i)_{i\in[s]}$ are independently and identically distributed (iid) with a normal distribution, as well as the entries of the $m \times r$ matrix K. For every considered instance, we use SeDuMi to solve the SOCP (5.7) and the $A-$optimality SDP (3.11); we have implemented the other procedures in Matlab. In all our experiments, the stopping criterion is based on the general equivalence theorem of Kiefer [Kie74]: the

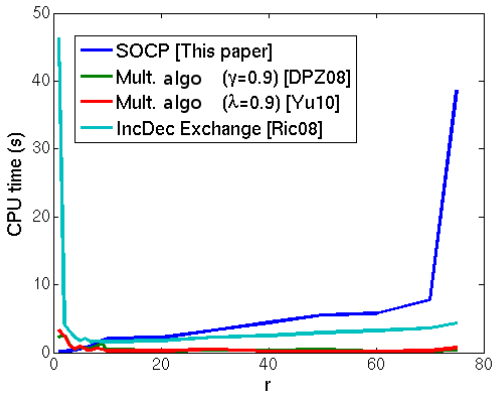


Figure 6.2: Comparison of four algorithms on typical random instances ($A-$optimality) with $m = 75$, $s = 150$, $l = 1$ and varying r.

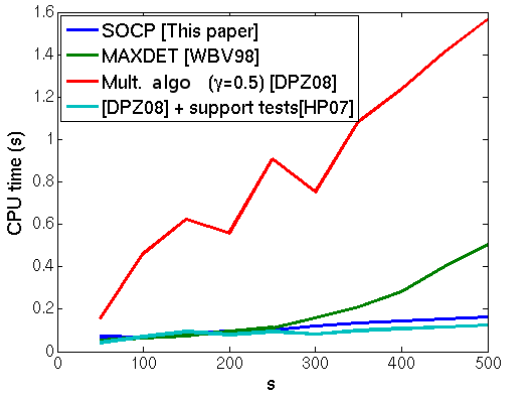


Figure 6.3: Comparison of four algorithms on typical random instances of the minimum covering ellipsoid ($D-$optimality for $\boldsymbol{\theta}$, $m = 3$) and varying s.

computation stops as soon as the ratio between the largest entry of the gradient and the value of the criterion is below 1.001 (as in [DPZ08]).

We start by evaluating the effect of r, which turns out to be the determining factor for the performance of our SOCP approach. To this end, we set $m = 75$, $s = 150$, $l = 1$ (single-response experiments), and we let r vary between 1 and 75. The computing time of the different algorithms is plotted against r in Figure 6.2. We notice that our algorithm is the fastest for small values ($r \leq 7$), but performs badly when r is large, while the multiplicative update algorithms are insensitive to the value of r. For this reason, we will chose small values of r in further experiments, since our algorithm might not be well adapted for large r.

We next study the effect of s (the number of available experiments) for the case of $\boldsymbol{c}-$optimality ($r = 1$). For these experiments, we set $m = 120$, $l = 30$, and we take s in the set $\{2^k, k = 2, \ldots, 11\}$. The performance (in terms of CPU time) of the SOCP is compared to that of the multiplicative algorithm with an acceleration parameter $\gamma = 0.9$ [DPZ08] on the log-log plot of Figure 6.1. The boxes represent the distribution of the CPU time, on 10 randomly generated instances. We see here that our approach is in average ten times faster as soon as $s \geq 32$.

To evaluate the effect of m, we set $s = 2^{10}$, $l = 1$, $r = 3$, and choose m in the set $\{2^k, k = 1, \ldots, 10\}$. (Note that since K and the A_i have random iid Gaussian entries, the instance is almost surely feasible if $s \geq m$; otherwise, the instance is almost surely infeasible.) The results of each algorithm are displayed in Table 6.1. It is striking that the SOCP approach is the best one, while the SDP is the worst when m becomes large, which demonstrates the importance of the rank reduction discussed in Chapter 4. For $m \leq 2^9$, the SOCP is 10 times faster than all other algorithms. In the last row of the table however, this ratio is lower. This might be because $s = m = 2^{10}$ in this case, such that all experiments are support points of the optimal design, and classic algorithms certainly take advantage of

this situation (while it does not make a difference for interior point codes).

Pronzato [Pro03] has shown that we can improve the multiplicative algorithms thanks to a simple test which allows to remove *on the fly* experiments which do not belong to the support of the $D-$optimal design (i.e. experiments with a zero weight), and which was refined by Harman and Pronzato [HP07]. This can considerably improve the performance of the multiplicative algorithms when there are a lot of points with a zero weight. As in [HP07], we have studied random instances of the minimum covering ellipse, but in $\mathbb{R}^3$: $m = 3$, $K = \boldsymbol{I}_3$, and we draw s independent random regression vectors ($l = 1$) from a normal distribution $\boldsymbol{a_i} \sim \mathcal{N}(0, \boldsymbol{I}_3)$, with s increasing from 50 to 500. The $D-$optimal design problem is equivalent to finding the minimum volume ellipsoid which contains the s vectors $\boldsymbol{a}_i$, and the $D-$optimal design is supported by points lying on the boundary of this minimal ellipsoid (Figure 3.1). In accordance with intuition, the number of support points of the $D-$optimal design is small, and therefore the test of Pronzato and Harman improves dramatically the computing time (cf. Figure 6.3). Note however that our SOCP for $D-$optimality (5.25) remains competitive with the latter approach.

6.2 Polynomial Regression

We have computed the $A-$ and D-optimal designs (for the full parameter $\boldsymbol{\theta}$), for a polynomial regression model of degree 5:

$$A(\boldsymbol{x}) = [1, \boldsymbol{x}, \boldsymbol{x}^2, \boldsymbol{x}^3, \boldsymbol{x}^4, \boldsymbol{x}^5]$$

on the regression region $\mathcal{X} = [0, 3]$. The optimal designs are represented on Figure 6.4. In this problem, we have $r = m = 6$, which is *small*. Therefore, we can hope that our SOCP approach will perform well. The computation times are plotted on Figures 6.5 and 6.6, as a function of the number of points considered for the discretization of the regression interval $\mathcal{X} = [0, 3]$. For the $A-$optimal design, the experimental setting was the same that the one of previous section. For the $D-$optimal design problem, we solved the geometric program (5.4) with SeDuMi. We have also implemented the classic multiplicative algorithm, the accelerated algorithm with $\gamma = 0.5$, and the MAXDET program (3.9). Contrarily to the multiplicative algorithms, the SOCP and the MAXDET approaches seem to be insensitive to the size of the discretization grid. For these instances, our SOCP is roughly two times faster than the MAXDET program. Also note that the effect of the acceleration parameter γ is clearly visible (red curve vs. green curve). We point out that for these polynomial regression problems, the tests of Pronzato and Harman [Pro03, HP07] to remove points that do not belong to the support of the $D-$optimal design did not yield any improvement.

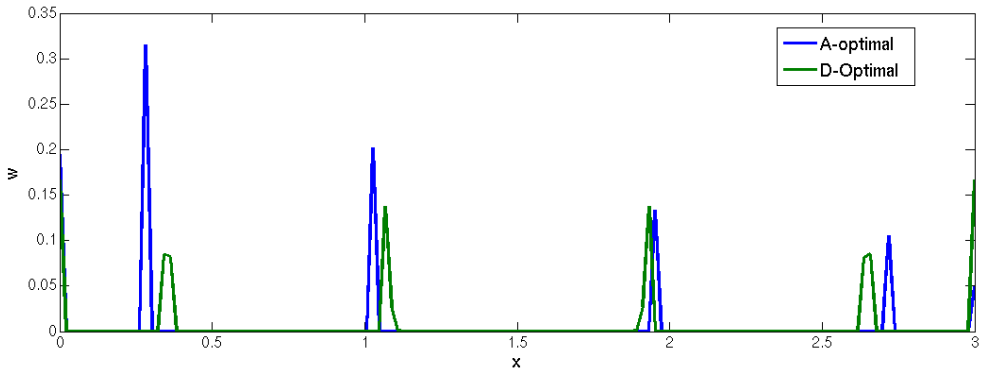


Figure 6.4: A- and D-optimal designs for the polynomial regression model of degree 5 on $\mathcal{X} = [0, 3]$.

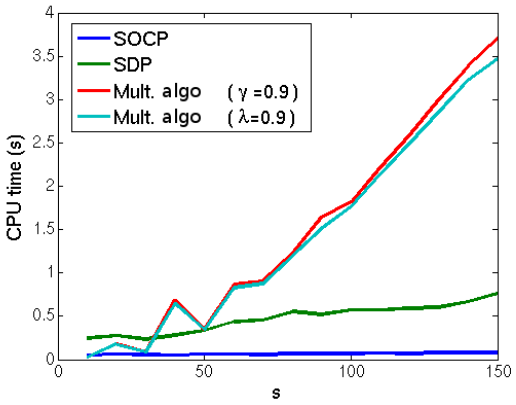


Figure 6.5: A-optimal design for the polynomial regression model: evolution of the computation time with the number of points for the discretization of $[0, 3]$.

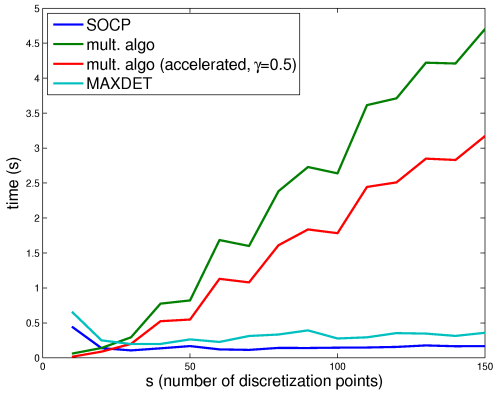


Figure 6.6: D-optimal design for the polynomial regression model: evolution of the computation time with the number of points for the discretization of $[0, 3]$.

6.3 Optimal Sampling in IP networks

We finally show some results for an application to the optimal monitoring of large IP networks. Assume that an Internet provider wants to estimate the traffic matrix of her network, that is, the volume of traffic between each pair of origin and destination during a given time period. To this end, she disposes of a monitoring tool, which can be activated at different sampling rates in different location of the network, and is able to find the destination of the sampled packets. For networking issues, the intensive use of this tool is not suitable, because it creates an overload both in terms of CPU utilization of the router and bandwidth consumption. The sampling rates should therefore be tuned cautiously on each interface, in such a way that the number of sampled packets remains under a target threshold.

This situation can be represented by an optimal design model with multiresponse experiments: the set of available experiments $\mathcal{X}$ coincides with the interfaces of the network where the monitoring-tool can be activated: when the software is installed on a given interface, we obtain an estimation of the sum of the flows that traverse this interface, and that have destination D, for every destination D reachable from this interface. In Chapter 10, we shall see that if the sampling rates are small, then the Fisher information matrix of the sampling design has the standard form (2.8) (after an appropriate normalization of the observation matrices relying on a prior estimate of the unknown OD traffic matrix). The optimal monitoring problem can thus be formulated as an optimal experimental design problem with multiple resource constraints.

We first study some $\boldsymbol{c}-$optimal sampling problem with the simple constraint $\sum_{i=1}^{s} w_i = 1$, such that we can compare our approach to classic algorithms. Table 6.2 summarizes the results (in terms of CPU time) for several problems: each instance is defined by a network and the type of interfaces considered. We used the topology of three networks: Abilene, which consists in 11 nodes, $m = 121$ OD pairs and 50 links; the Opentransit backbone of France Telecom, with 116 nodes, $m = 13456$ OD pairs and 436 links; and a clustered version of the latter network, thus reduced to 31 nodes, $m = 961$ OD pairs and 133 links. The natural problem is to activate the monitoring tool independently on each link (interfaces="links"). We also considered the problem of imposing the same sampling rates on all incoming links of each router, which is equivalent to consider each router as a *big interface* (interfaces="Nodes"). This setting is consistent with older versions of the monitoring software Netflow, still present on many routers in practice, and which do not allow to set different sampling rates on different incoming interfaces. For all these instances the vector $\boldsymbol{c}$ was drawn from a normal distribution. The threshold for the stopping criterion was lowered to 1.01 for this network application, since this value suffices to obtain good designs in practice.

Network	Abilene ($m = 121$)	Abilene ($m = 121$)	OTClusters ($m = 961$)	OTClusters ($m = 961$)	Opentransit ($m = 13456$)	Opentransit ($m = 13456$)
Interfaces	Nodes ($s = 11$)	Links ($s = 50$)	Nodes ($s = 31$)	Links ($s = 133$)	Nodes ($s = 116$)	Links ($s = 436$)
SOCP	0.021	0.036	0.078	0.094	5.52	33.03
SDP	1.095	1.178	692.37	734.25	failed	failed
IncDec Exchange	0.518	0.823	4.57	19.69	failed	failed
Mult. algo ($\gamma = 0.9$)	0.009	0.043	0.018	1.893	failed	failed
Mult. algo ($\lambda = 0.9$)	0.008	0.038	0.018	1.468	failed	failed

Table 6.2: CPU time (s) for different instances of $\boldsymbol{c}-$optimal design arising from an optimal monitoring problem in IP networks (with the standard constraint $\sum_i w_i = 1$)

We can see in the table that the multiplicative algorithms perform better than the SOCP approach on the instances where s is small (1st and 3rd columns in Table 6.2). On the other instances however, the SOCP performs well, and it is the only method which returned a solution for the Opentransit network. The SDP and the multiplicative methods failed because of memory issues (in the multiplicative algorithm, a full rank update of the

13456×13456 information matrix should be carried out at each time step). The IncDec Exchange algorithm did not crash, but it had not converged after 2 hours of computation.

We next turn to the case of general constraints of the form $R\boldsymbol{w} \leq \boldsymbol{d}$. Since we do not know any other algorithm which can handle optimal design problems with multiple resource constraints, we compare the SOCP and the semi-definite programming approaches only. Table 6.3 summarize the results (in terms of CPU time) for several problems, specified as previously by the network and the type of interfaces considered, and also by the type of the constraint matrix R. In the optimal sampling problem, the matrix R usually depends on the volume of traffic observed at each router (cf. [SGB10]). We simulated this data from a uniform distribution, a lognormal distribution, or we used real traffic loads. To see the effect of the number of constraints, we also generated arbitrary constraints matrices of different sizes.

In comparison to the SDP, the computation time can be reduced by a factor in the order of 1000 on the instances from the clustered network. Moreover, the SOCP approach is able to handle huge instances arising from the Opentransit network (in which $m = 13456$).

Network	Abilene ($m = 121$)	Abilene ($m = 121$)	Abilene ($m = 121$)	Abilene ($m = 121$)	Abilene ($m = 121$)
Interfaces	Links ($s = 50$)	Links ($s = 50$)	Links ($s = 50$)	Nodes ($s = 11$)	Nodes ($s = 11$)
Constraints	R: 11×50 (uniform traffic)	R: 11×50 (lognormal traffic)	R: 11×50 (real traffic)	R: 4×11 (arbitrary)	R: 10×11 (arbitrary)
SOCP	0.043	0.056	0.061	0.051	0.053
SDP	0.714	0.842	0.944	0.827	0.876

Network	OTClusters ($m = 961$)	OTClusters ($m = 961$)	OTClusters ($m = 961$)	Opentransit ($m = 13456$)	Opentransit ($m = 13456$)
Interfaces	Nodes ($s = 31$)	Links ($s = 133$)	Links ($s = 133$)	Links ($s = 436$)	Links ($s = 436$)
Constraints	R: 4×31 (arbitrary)	R: 31×133 (uniform traffic)	R: 130×133 (arbitrary)	R: 12×436 (arbitrary)	R: 116×436 (real traffic)
SOCP	0.141	0.462	1.135	23.32	187.59
SDP	350.63	451.69	430.71	failed	failed

Table 6.3: Computation time (s) for different instances of $\boldsymbol{c}-$optimal design arising from an optimal monitoring problem in IP networks (with multiple constraints $R\boldsymbol{w} \leq \boldsymbol{b}$).

Chapter 7

Combinatorial problems arising in optimal design of experiments

In this chapter, we study some combinatorial aspects of optimal experimental design problems. The results of this chapter are presented in [Sag10]. Some of them were already announced in [BGS08].

In a number of real-world applications, the design variables are discrete, since the experimenter can only choose the experiments to conduce from a finite set, and perhaps how many times to perform them. An exhaustive list of these applications is not possible, but we wish to give the reader a few examples from these problems:

Uciński and Patan [UP07] interested themselves in the estimation of parameters of systems governed by partial differential equations. They propose to solve a D-optimal problem in order to find an optimal subset of spatial locations of sensors on a finite grid. Their approach is based on a Branch and Bound algorithm, where a multiplicative algorithm is used to solve a continuous relaxation of the problem and provides some upper bounds.

Song, Qiu and Zhang [SQZ06] proposed an application of the optimal experimental design for the estimation of performance in a large scale network. In their approach, a discrete A- (or D-)optimal design is approximated by a greedy algorithm in order to select some measurements of the network performance. This greedy algorithm entails smart rank-one matrix updates, as first suggested by Fedorov [Fed72].

Branderhorst, Walmsley, Kosut and Rabitz [BWKR08] used the optimal design framework to maximize the accuracy of the estimation of quantum states, by selecting the number of experiments to be performed in each particular system configuration. A continuous relaxation of the problem is solved, and they rounded to obtain an integer solution.

Finally, the present developments were motivated by a joint work with Bouhtou and Gaubert [BGS08, SGB10] (see also Chapter 10) on the application of optimal experiment design methods to the identification of the traffic on an Internet backbone. The approach developed there consists in solving the continuous relaxation of an optimal experimental

design problem, which is rounded with a simple heuristic in order to obtain a discrete design.

The rest of this chapter is organized as follows: in Section 7.1 we introduce the notation and we state the combinatorial optimization problem which we shall study; particular care will be given to the under-instrumented situation, where no discrete design lets the information matrix be of full rank, and which may occur in monitoring problems on large size networks. To the best of our knowledge, this combinatorial optimization problem has always been handled by heuristic approaches. This chapter provides approximability bounds for this NP-hard problem.

In Section 7.2, we show that this combinatorial optimization problem is NP-hard indeed, and we establish a matrix inequality (Proposition 7.2.4) which shows that a class of spectral functions is submodular (Corollary 7.2.5). As a particular case of the latter result, the objective function of the experimental design problem is submodular. Due to a celebrated result of Nemhauser, Wolsey and Fisher [NWF78], this implies that the greedy approach, which has often been used for this problem, always gives a design within $1 - e^{-1}$ of the optimum (Theorem 7.2.7). We point out that the submodularity of the D-criterion was noticed earlier: Robertazzi and Schwartz used it to write an accelerated Wynn-Fedorov–type algorithm for the construction of approximate designs [RS89] (i.e. with the constraint $\sum_i w_i = 1, \mathbf{0} \leq \boldsymbol{w} \leq \mathbf{1}$), which is based on the accelerated greedy algorithm of Minoux [Min78]. The originality of this chapter is to show that a whole class of criteria satisfies the submodularity property, and to study the consequences in terms of approximability of a combinatorial optimization problem.

In Section 7.3, we study some rounding algorithms for the optimal experimental design. Rounding a continuous solution to obtain a discrete one is a natural option [BWKR08, BGS08] since we dispose of a continuous relaxation of the problem, which is convex and has been extensively studied. Moreover, we may exploit the work of Calinescu, Chekuri, Pál and Vondrák [CCPa07, Von08], who showed how to use the pipage rounding algorithm of Ageev and Sviridenko [AS04] to approximate the maximization of submodular functions. Thanks to their ideas indeed, we show in Theorem 7.3.7 that when the goal is to select n out of s experiments, the $D-$optimal design may be rounded to a design for which the dimension of the observable subspace is within $\frac{n}{s}$ of the optimum. While this result might look weaker than the greedy $(1 - e^{-1})-$approximation factor, we show that one can not hope for a better result with rounding algorithms. The proof is based on a generalization of a result from Atwood [Atw73], who showed that the coordinates of the $D-$optimal design for experiments with scalar response are bounded by $\frac{1}{m}$, where m is the number of parameters to estimate. For multiresponse experiments, we generalized his result in Proposition 7.3.4, with inequalities involving the ranks of the observation matrices.

7.1 Notation and statement of the problem

7.1.1 A combinatorial optimization problem

We consider the same model as the one described in Chapter 2 (Each experiment provides linear multidimensional observations of the parameter (cf. Equation with a unit, centered noise (2.3).) In addition, we dispose of a prior observation

$$\boldsymbol{y_0} = A_0\boldsymbol{\theta} + \boldsymbol{\epsilon_0}.$$

We use the index 0 to denote this prior information. This can be useful to model a *free-of-charge* experiment, that the experimenter will conduct in any case, or to model an intrinsic relationship between the parameters, such as Kirchhoff's circuit law (cf. Section 5.2.3).

In this chapter, we assume that the experimenter wants to choose a well suited subset $\mathcal{I} \subseteq [s]$ of experiments that she will conduct in order to estimate the parameters. We therefore define the *design* variable $\boldsymbol{w}$ as the $0/1$ vector of size s, where w_k takes the value 1 if and only if $k \in \mathcal{I}$. We denote by $\mathcal{I} = \{i_1, \dots, i_n\}$ the subset of the selected experiments, such that the vector of observation reads :

$$\boldsymbol{y} = A(\boldsymbol{w})\ \boldsymbol{\theta} + \boldsymbol{\epsilon}, \tag{7.1}$$

$$\text{where } \boldsymbol{y} = \begin{bmatrix} \boldsymbol{y_0} \\ \boldsymbol{y_{i_1}} \\ \vdots \\ \boldsymbol{y_{i_n}} \end{bmatrix}, \quad A(\boldsymbol{w}) = \begin{bmatrix} A_0 \\ A_{i_1} \\ \vdots \\ A_{i_n} \end{bmatrix}, \quad \text{and} \quad \mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0},\ \mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) = I.$$

If we have enough measurements, such that $A(\boldsymbol{w})$ is of full rank, then $M(\boldsymbol{w}) = A(\boldsymbol{w})^T A(\boldsymbol{w}) = \sum_{i=1}^{s} w_i A_i^T A_i$ is the inverse of the covariance matrix for the best linear unbiased estimator of $\boldsymbol{\theta}$ (cf. Chapter 2). We can thus formulate the Φ_p-optimization problem in the same form as the one presented in Section 2.3.2, except that the design variable $\boldsymbol{w}$ is now integer, and subject to a cardinality constraint:

$$\max_{\boldsymbol{w}\ \in\{0,1\}^s} \Phi_p\Big(M(\boldsymbol{w})\Big) \tag{7.2}$$
$$s.t. \quad \sum_{i=1}^{s} w_i \leq n$$

Assume more generally that the cost of experiment i is r_i. If the experimenter has a

limited budget b, the (combinatorial) Φ_p-*optimal design* problem is:

$$\max_{\boldsymbol{w} \in\{0,1\}^s} \Phi_p\Big(M(\boldsymbol{w})\Big) \qquad (7.3)$$
$$s.t. \quad \sum_{i=1}^{s} w_i r_i \leq b$$

Problem (7.2) is a particular case of Problem (7.3), when all the experiments have the same cost $\overline{r}$, and $n = \lfloor \frac{b}{\overline{r}} \rfloor$. Therefore, we refer to the constraints of Problem (7.2) as the *unit-cost* case.

7.1.2 The under-instrumented situation

We note that the problem of maximizing $M(\boldsymbol{w})$ with respect to the Löwner ordering remains meaningful even when $M(\boldsymbol{w})$ is not of full rank. This case does arise in under-instrumented situations, in which some constraints may not allow one to conduct a number of experiments which is sufficient to infer all the parameters. In this case however, the natural interpretation of $M(\boldsymbol{w})$ as *the inverse of the covariance matrix of the best linear unbiased estimator* vanishes, because an unbiased estimator for the vector of parameters does not exist. In a number of applications though, the parameters can still be estimated, using a small number of measurements and prior information on $\boldsymbol{\theta}$. Therefore, a measure of the quality of the under-instrumented designs is required.

An interesting and natural idea to find an optimal under-instrumented design is to choose the design which maximizes the rank of the observation matrix $A(\boldsymbol{w})$, or equivalently of $M(\boldsymbol{w}) = A(\boldsymbol{w})^T A(\boldsymbol{w})$. The *rank maximization* is a nice combinatorial problem, where we are looking for a subset of matrices whose sum is of maximal rank:

$$\max_{\boldsymbol{w}\in\{0;1\}^s} \operatorname{rank}\left(A_0^T A_0 + \sum_i w_i A_i^T A_i\right) - \operatorname{rank}(A_0^T A_0) \qquad (P_0)$$
$$s.t. \qquad \sum_i w_i r_i \leq b.$$

In the above optimization problem, the term $\operatorname{rank}(A_0^T A_0)$ has been subtracted so that the objective criterion takes the value 0 for $\boldsymbol{w} = \mathbf{0}$. In combinatorics, approximation factors are generally given with respect to objective functions which satisfy the latter property.

More generally, we show below that the problem of maximizing $M(\boldsymbol{w})$ with respect to the Löwner ordering still has some statistical interest in the under-instrumented situation. Moreover, we will see that the Φ_p-maximization of $M(\boldsymbol{w})$ may be thought as a regularization of the *rank optimization* problem (P_0), and Φ_p can be seen as a deformation of the rank criterion for $p \in]0, 1]$. First, we show that $M(\boldsymbol{w})$ still has a statistical meaning, since its Moore-Penrose generalized inverse is the variance of the estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{LS}}$ given by least square theory. More precisely, a linear estimator $\hat{\boldsymbol{\theta}} = L^T \boldsymbol{y}$ for $\boldsymbol{\theta}$ is unbiased if and

only if L^T is a left inverse of $A(\boldsymbol{w})$ (i.e. $L^T A(\boldsymbol{w}) = I$). In the under-observed case, no such left inverse exists, but we know from least square theory that the trace of the covariance matrix $\mathrm{Var}(\hat{\boldsymbol{\theta}}) = L^T L$ is minimized in the class of the least biased estimators for $L^* = (A(\boldsymbol{w})^T)^\dagger$, where $M^\dagger$ denotes the Moore-Penrose generalized inverse of M (i.e. L^* minimizes $\|L\|_F := \operatorname{trace} L^T L$ in the class of matrices L such that $\|(L^T A(\boldsymbol{w}) - I)\|_F$ is minimized). The resulting least square estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{LS}} = A(\boldsymbol{w})^\dagger \boldsymbol{y}$ has variance

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}_{\boldsymbol{LS}}) = A(\boldsymbol{w})^\dagger (A(\boldsymbol{w})^\dagger)^T = \Big(A(\boldsymbol{w})^T A(\boldsymbol{w})\Big)^\dagger = M(\boldsymbol{w})^\dagger.$$

Similarly to the full rank case (cf. Equation 2.11), we can see that for all $\alpha \in [0,1]$, $\hat{\boldsymbol{\theta}}_{\boldsymbol{LS}}$ lies in a cylinder of the form

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{LS}})^T M(\boldsymbol{w}) (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{LS}}) \leq \kappa_\alpha$$

with probability α, and these cylinders are minimized (for the inclusion relation) when $M(\boldsymbol{w})$ is maximized (for the Löwner ordering).

Another argument for the use of this criterion is given by Bayesian optimal design, where a prior distribution for the parameter is given:

$$\mathbb{E}(\boldsymbol{\theta}) = \boldsymbol{\mu}, \quad \text{and} \quad \mathrm{Var}(\boldsymbol{\theta}) = R.$$

It is known (see e.g. [Puk93]) that when the prior covariance matrix R is positive definite, the expected covariance matrix is minimized among all unbiased affine estimators, conditionally to the prior distribution of $\boldsymbol{\theta}$ for:

$$\hat{\boldsymbol{\theta}}_{|R,\boldsymbol{\mu}} = \Big(R^{-1} + A(\boldsymbol{w})^T A(\boldsymbol{w})\Big)^{-1} \Big(R^{-1}\boldsymbol{\mu} + A(\boldsymbol{w})^T \boldsymbol{y}\Big).$$

This Bayesian estimator has a variance which does not depend on the prior expected value of $\boldsymbol{\theta}$:

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}_{|R,\mu}) = \Big(R^{-1} + A(\boldsymbol{w})^T A(\boldsymbol{w})\Big)^{-1}. \tag{7.4}$$

In practice, prior information on the variance of the parameter is rarely known, and the prior can be modeled instead by setting $R^{-1} = \epsilon I$ for an arbitrarily small ϵ (see e.g. [SQZ06]). The regularization parameter ϵ lets the inverse in (7.4) exist, and we recover the Moore-Penrose inverse of $M(\boldsymbol{w})$ by letting $\epsilon \to 0$.

When every feasible information matrix is singular, Equation (2.13) indicates that the maximization of $\Phi_p(M(\boldsymbol{w}))$ can be considered only for nonnegative values of p. The next proposition shows that Φ_p can be seen as a deformation of the rank criterion for $p \in]0,1]$.

First notice that when $p > 0$, the maximization of $\Phi_p(M(\boldsymbol{w}))$ is equivalent to:

$$\begin{aligned} \max_{\boldsymbol{w}\in\{0,1\}^s} \quad & \varphi_p(\boldsymbol{w}) := \ \mathrm{trace}\left(A_0^T A_0 + \sum_k w_k A_k^T A_k\right)^p - \mathrm{trace}\left(A_0^T A_0\right)^p \\ s.t. \quad & \sum_k w_k c_k \leq b, \end{aligned} \tag{P_p}$$

where we have subtracted the term $\mathrm{trace}(A_0^T A_0)^p$ from the objective function, as in Problem (P_0), in order to have the property $\varphi_p(\mathbf{0}) = 0$.

Proposition 7.1.1. *For all positive semidefinite matrix $M \in \mathbb{S}_m^+$,*

$$\lim_{p\to 0^+} \mathrm{trace}\ M^p = \mathrm{rank}\ M. \tag{7.5}$$

Proof. Let $\lambda_1, \ldots, \lambda_r$ denote the positive eigenvalues of M, counted with multiplicities, such that r is the rank of M. We have the first order expansion as $p \to 0^+$:

$$\mathrm{trace}\ M^p = \sum_{k=1}^{r} \lambda_k^p = r + p\ \log(\prod_{k=1}^{r} \lambda_k) + \mathcal{O}(p^2) \tag{7.6}$$

□

Consequently, $\mathrm{trace}\ M^0$ will stand for $\mathrm{rank}(M)$ in the sequel and the rank maximization problem (P_0) is the limit of problem (P_p) as $p \to 0^+$.

Corollary 7.1.2. *If $p > 0$ is small enough, then every design $\boldsymbol{w}^*$ which is a solution of Problem (P_p) maximizes the rank of $M(\boldsymbol{w})$. Moreover, among the designs which maximize this rank, $\boldsymbol{w}^*$ maximizes the product of nonzero eigenvalues of $M(\boldsymbol{w})$.*

Proof. Since there is only a finite number of designs, it follows from (7.6) that for $p > 0$ small enough, every design which maximizes φ_p must maximize in the lexicographical order first the rank of $M(\boldsymbol{w})$, and then the product $\prod_{\lambda_k>0} \lambda_k$. □

7.2 Submodularity and Greedy approach

In this section, we study the greedy algorithm for solving Problem (P_p) through the submodularity of φ_p. We will first prove that the *rank optimization* problem is NP-hard by reduction of MAX-k-Coverage. Next, we show that the objective function of Problem (P_p) is *nondecreasing submodular*. The maximization of submodular functions over a matroid has been extensively studied [NWF78, CCPa07, Von08, KST09], and we shall use known approximability results.

In combinatorics, approximability results are usually given for optimization problems whose objective function takes the value 0 for the empty set. For this reason, all results will

be given with respect to the maximization of the function φ_p (Problem (P_p)). This problem is equivalent to the Φ_p-optimal problem (7.3) for positive values of p, and to the *rank optimization* problem (P_0) for $p = 0$. In addition, note that there is no point to consider multiplicative approximation factors for the Φ_p-optimal problem when $p \leq 0$, since the criterion is identically 0 as long as the the information matrix is singular. For $p \leq 0$ indeed, the instances of the Φ_p-optimal problem where no feasible design lets $M(\boldsymbol{w})$ be of full rank have an optimal value of 0. For all the other instances, any polynomial-time algorithm with a positive approximation factor would necessarily return a design of full rank. Provided that $P \neq NP$, this would contradict the NP-hardness of the rank optimization problem (Theorem 7.2.1). So, we investigate approximation algorithms only in the case $p \geq 0$.

7.2.1 Hardness of *Rank optimization*

Theorem 7.2.1. *Problem (P_0) is NP-Hard. For all positive ε, there is no polynomial-time algorithm which approximates (P_0) by a factor of $1 - \frac{1}{e} + \varepsilon$ unless $P = NP$.*

Proof. We will show that the problem MAX-k-coverage, for which the statement of the theorem is true [Fei98], reduces to the *rank optimization* (P_0) in polynomial time.

The problem MAX-k-Coverage is defined as follows : We are given a collection of subsets $\mathcal{S} = \{S_1, S_2, \ldots, S_s\}$ of $[m]$, as well as an integer k, and the goal is to pick at most k sets of $\mathcal{S}$ such that the size of their union is maximized. Let $\boldsymbol{e_i}$ be the i^{th} vector of the canonical basis of $\mathbb{R}^m$. If the set S_i contains the elements $\{i_1, i_2, \ldots, i_{l(i)}\}$, we define the i^{th} observation matrix as: $A_i = [\boldsymbol{e_{i_1}}, \ldots, \boldsymbol{e_{i_{l(i)}}}]^T$, such that $A_i^T A_i$ is the diagonal matrix whose indexes of nonzero entries are the elements of S_i. Finally, let A_0 be the all-zero row vector of size m. Since all the matrices $A_i^T A_i$ have only diagonal entries, it is straightforward to see that the rank of $A_0^T A_0 + \sum_k w_k A_k^T A_k$ is equal to the number of nonzero elements on its diagonal, i.e. the cardinal of $\cup_{\{i|w_i=1\}} S_i$, which is exactly the objective function of the MAX-k-Coverage problem. □

This is a negative result on the approximability of Problem (P_p). Nevertheless, we show that the bound provided by Theorem 7.2.1 is the worst possible ever, and that the greedy algorithm always attains it in the unit-cost case.

7.2.2 A class of submodular spectral functions

We recall that a real valued function $F : 2^E \to \mathbb{R}$, defined on every subset of E is called nondecreasing if for all subsets I and J of E, $I \subseteq J$ implies $F(I) \leq F(J)$. We also give the definition of a *submodular* function:

Definition 7.2.2 (Submodularity)**.** A real valued set function $F : 2^E \longrightarrow \mathbb{R}$ is *submodular* if it satisfies the following conditions :

(i) $F(\emptyset) = 0$;

(ii) $F(I) + F(J) \geq F(I \cup J) + F(I \cap J)$ for all $I, J \subseteq E$.

We next recall the definition of operator monotone functions. The latter are real valued functions applied to Hermitian matrices: if $A = U\,\mathrm{Diag}(\lambda_1, \ldots, \lambda_m)U^*$ is a $m \times m$ Hermitian matrix (where U is unitary and U^* is the conjugate of U), the matrix $f(A)$ is defined as $U\,\mathrm{Diag}(f(\lambda_1), \ldots, f(\lambda_m))U^*$.

Definition 7.2.3 (Operator monotonicity)**.** A real valued function f is *operator monotone* on $\mathbb{R}_+$ (resp. $\mathbb{R}_+^*$) if for every pair of positive semidefinite (resp. positive definite) matrices A and B

$$A \preceq B \Longrightarrow f(A) \preceq f(B).$$

We say that f is *operator antitone* if $-f$ is operator monotone.

The next proposition is a matrix inequality of independent interest; it will be useful to show that φ_p is submodular. Interestingly, it can be seen as an extension of the Ando-Zhan Theorem [AZ99], which reads as follows: *Let A, B be positive semidefinite matrices. For any unitarily invariant norm $|||\cdot|||$, and for every nonnegative operator monotone function f on $[0, \infty)$,*

$$|||f(A + B)||| \leq |||f(A) + f(B)|||.$$

Kosem [Kos06] asked whether it is possible to extend this inequality as follows:

$$|||f(A + B + C)||| \leq |||f(A + B) + f(B + C) - f(C)|||,$$

and gave a counter example involving the trace norm and the function $f(x) = \frac{x}{x+1}$. However, we show in next proposition that the previous inequality holds for the trace norm and every primitive f of an operator antitone function (in particular, for $f(x) = x^p,\ p \in]0, 1]$). Note that the previous inequality is not true for any unitarily invariant norm and $f(x) = x^p$ neither. It is easy to find counter examples with the spectral radius norm.

Proposition 7.2.4. *Let f be a real function defined on $\mathbb{R}_+$ and differentiable on $\mathbb{R}_+^*$. If f' is operator antitone on $\mathbb{R}_+^*$, then for all triple (X, Y, Z) of $m \times m$ positive semidefinite matrices,*

$$\mathrm{trace}\ f(X + Y + Z) + \mathrm{trace}\ f(Z) \leq \mathrm{trace}\ f(X + Z) + \mathrm{trace}\ f(Y + Z). \tag{7.7}$$

Proof. Since the eigenvalues of a matrix are continuous functions of its entries, and since $\mathbb{S}_m^{++}$ is dense in $\mathbb{S}_m^+$, it suffices to establish the inequality when X,Y, and Z are positive definite. Let X be an arbitrary positive definite. We consider the map:

$$\begin{aligned} \psi : \mathbb{S}_m^+ &\longrightarrow \mathbb{R} \\ T &\longmapsto \mathrm{trace}\ f(X + T) - \mathrm{trace}\ f(T). \end{aligned}$$

The inequality to be proved can be rewritten as

$$\psi(Y+Z) \leq \psi(Z).$$

We will prove this by showing that ψ is nonincreasing with respect to the Löwner ordering in the direction generated by any positive semidefinite matrix. To this end, we compute the Frechet derivative of ψ at $T \in \mathbb{S}_m^{++}$ in the direction of an arbitrary matrix $H \in \mathbb{S}_m^{+}$. By definition,

$$D\psi(T)[H] = \lim_{\epsilon \to 0} \frac{1}{\epsilon}\Big(\psi(T+\epsilon H) - \psi(T)\Big).$$

When f is an analytic function, $X \longmapsto \operatorname{trace} f(X)$ is Frechet-differentiable, and an explicit form of the derivative is known (see [HP95, JB06]): $D\Big(\operatorname{trace} f(A)\Big)[B] = \operatorname{trace}\Big(f'(A)B\Big)$. Since f' is operator antitone on $\mathbb{R}_+^*$, a famous result of Löwner [Löw34] tells us (in particular) that f' is analytic at all point of the positive real axis, and the same holds for f. Provided that the matrix T is positive definite (and hence $X+T \succ 0$), we have

$$D\psi(T)[H] = \operatorname{trace}\Big(\Big(f'(X+T) - f'(T)\Big)H\Big).$$

By antitonicity of f' we know that the matrix $W = f'(X+T) - f'(T)$ is negative semidefinite. For a matrix $H \succeq 0$, we have therefore:

$$D\psi(T)[H] = \operatorname{trace}\,(WH) \leq 0.$$

Consider now $h(s) := \psi(sY+Z)$. For all $s \in [0,1]$, we have

$$h'(s) = D\psi(sY+Z)[Y] \leq 0,$$

and so, $h(1) = \psi(Y+Z) \leq h(0) = \psi(Z)$, from which the desired inequality follows. □

Corollary 7.2.5. *Let $M_0, M_1, \ldots, M_s$ be $m \times m$ positive semidefinite matrices. If f satisfies the assumptions of Proposition 7.2.4, then the set function $F : 2^{[s]} \to \mathbb{R}$ defined by*

$$\forall I \subset [s],\ F(I) = \operatorname{trace}\ f(M_0 + \sum_{i \in I} M_i) - \operatorname{trace}\ f(M_0),$$

is submodular

Proof. The relation $F(\emptyset) = 0$ follows from the definition of F.

Let $I, J \subseteq 2^{[s]}$. We define

$$X = \sum_{i \in I \setminus J} M_i,\ Y = \sum_{i \in J \setminus I} M_i,\ Z = A_0^T A_0 + \sum_{i \in I \cap J} M_i.$$

It is easy to check that

$$\begin{aligned} F(I) &= \text{trace } f(X+Z) - \text{trace } f(M_0), \\ F(J) &= \text{trace } f(Y+Z) - \text{trace } f(M_0), \\ F(I\cap J) &= \text{trace } f(Z) - \text{trace } f(M_0), \\ F(I\cup J) &= \text{trace } f(X+Y+Z) - \text{trace } f(M_0). \end{aligned}$$

Hence, Proposition 7.2.4 proves the submodularity of F. □

We next point out some submodular set functions which can be found thanks to Corollary 7.2.5.

Corollary 7.2.6. *Let $M_0, M_1, ..., M_s$ be $m \times m$ positive semidefinite matrices.*

(i) $\forall p \in]0,1], I \mapsto \text{trace}(M_0 + \sum_{i\in I} M_i)^p - \text{trace } M_0^p$ *is submodular.*

(ii) $I \mapsto \text{rank}(M_0 + \sum_{i\in I} M_i) - \text{rank } M_0$ *is submodular.*

If moreover M_0 is positive definite, or if every M_i is positive definite, then:

(iii) $I \mapsto \log\det(M_0 + \sum_{i\in I} M_i) - \log\det M_0$ *is submodular.*

Proof. It is known that $x \mapsto x^q$ is operator antitone on $\mathbb{R}_+^*$ for all $q \in [-1,0[$. Therefore, the derivative of the function $x \mapsto x^p$ (which is px^{p-1}), is operator antitone on $\mathbb{R}_+^*$ for all $p \in]0,1[$. This proves the point (i) for $p \neq 1$. The case $p = 1$ is trivial, by linearity of the trace.

The submodularity of the rank (ii) and of $\log\det$ (iii) are classic. Interestingly, they are obtained as the limit case of (i) as $p \to 0^+$. (For $\log\det$, we must consider the second term in the asymptotic development of $X \mapsto \text{trace } X^p$ as p tends to 0^+ (7.6)). □

7.2.3 Greedy approximation

The next results show that for all $p \in [0;1]$, Problem (P_p) is $1 - \frac{1}{e}-$approximable in polynomial time. This can be attained with the help of the greedy algorithm, whose principle is to start from $\mathcal{G}_0 = \emptyset$ and to construct sequentially the sets

$$\mathcal{G}_{k+1} := \mathcal{G}_k \cup \text{argmax}_{i\in[s]} \frac{\varphi_p(\mathcal{G}_k \cup \{i\})}{r_i},$$

until the budget constraint is violated.

Theorem 7.2.7 (Approximability of φ_p-Optimal Design: Unit-cost case)**.** *Let $p \in [0;1]$. The greedy algorithm for problem (P_p) yields a $1 - \frac{1}{e}$ approximation factor in the unit-cost case.*

Proof. We know from Corollary 7.2.6 that for all $p \in [0,1]$, φ_p is submodular ($p = 0$ corresponding to the rank maximization problem). In addition, the function φ_p is nondecreasing, because $X \longrightarrow X^p$ is a matrix monotone function for $p \in [0,1]$ (see e.g. [Zha02]).

Nemhauser, Wolsey and Fisher [NWF78] proved the result of this theorem for any nondecreasing submodular function over a uniform matroid. Moreover when the maximal number of interfaces which can be selected is n, this approximation ratio can be improved to $1 - \Big(1 - 1/n\Big)^n$. □

Remark 7.2.1. As mentionned in the introduction of this chapter, the submodularity of the $D-$criterion was already used by Robertazzi and Schwartz [RS89]. The problem studied in the latter article is of a different nature, since the authors used a greedy algorithm to solve Problem (7.2) (for $p = 0$) when $n \to \infty$, and they normalize the result to obtain an optimal approximate design. The submodularity of Φ_0 allowed them to use the accelerated greedy algorithm of Minoux [Min78]. This yields great computational savings, because at each stage, the increment of the objective function need only be computed for a subset of $[s]$. Note that this accelerated greedy algorithm can also be used in our case, in order to construct a $1 - 1/e-$approximation of the φ_p-optimum.

One can obtain a better bound by considering the *total curvature* of a given instance, which is defined by:

$$c = \max_{i \in [s]} \; 1 - \frac{\varphi_p\Big([s]\Big) - \varphi_p\Big([s] \setminus \{i\}\Big)}{\varphi_p\Big(\{i\}\Big)} \in [0,1].$$

Corollary 7.2.8 (Approximability of φ_p-Optimal Design in function of the curvature)**.** *Let $p \in [0;1]$, and c be the total curvature of a given instance of the Problem (P_p) in the unit-cost case, where the maximum number of experiments to be selected is n. The greedy algorithm for problem (P_p) yields a $\frac{1}{c}\Big(1 - (1 - \frac{c}{n})^n\Big)$ approximation factor.*

Proof. This result follows from Conforti and Cornuejols [CC84], who proved it for the maximization of an arbitrary nondecreasing submodular function with total curvature c. □

For the value $p = 1$, φ_p is additive and it follows that the total curvature is $c = 0$, yielding an approximation factor of 1, since

$$\lim_{c \to 0^+} \frac{1}{c}\Big(1 - (1 - \frac{c}{n})^n\Big) = 1.$$

As a consequence, the greedy algorithm always give the optimal solution of the problem. Note that Problem (P_1) is nothing but a *knapsack* problem, for which it is well known that the greedy algorithm is optimal in the unit-cost case. However, it is not possible to give a lower bound on the total curvature c for other values of $p \in [0,1[$, and c has to be computed for each instance. We now give a result for the budgeted problem:

Corollary 7.2.9 (Approximability of φ_p-Optimal Design)**.** *Problem* (P_p) *is still* $1-\frac{1}{e}-$*approximable in polynomial time in the budgeted case, but the greedy algorithm for problem* (P_p) *yields a constant approximation factor of only* $\frac{1}{2}\left(1-\frac{1}{e}\right)$.

Proof. This was proved for an arbitrary nondecreasing submodular function in [Svi04]. In order to attain the $1-1/e-$approximation guarantee, one can associate the greedy algorithm with the partial enumeration of all triples of experiments. □

Remark 7.2.2. The results of this section hold in particular for $p = 0$, and hence for the *rank maximization* problem (P_0).

7.3 Approximation by randomized rounding algorithms

The optimal design problem has a natural continuous convex relaxation which is simply obtained by removing the $0/1-$constraint on the design variable $\boldsymbol{w}$, and has been extensively studied (cf. Chapter 3). As mentioned in the introduction of this chapter, several authors proposed to solve this convex relaxation and to round the solution to obtain a near-optimal discrete design. We next investigate the legitimacy of this technique. We show in Theorem 7.3.7 that the D-optimal design may be rounded to a random discrete design which approximates the optimum of the rank optimization problem (P_0) by an average factor of $\frac{n}{s}$. While this result may look rather worse than the greedy approximation factor presented in Section 7.2, it is (almost) optimal since there are some instances for which the average ratio of approximation is $\frac{n}{s-1}$ (cf. Remark 7.3.2).

Another motivation for this section arises from the recent results from Calinescu, Chekuri, Pál and Vondrák [CCPa07, Von08], who showed that the problem of maximizing a nondecreasing submodular function over an arbitrary matroid is $(1-e^{-1})$-approximable, by first *approaching the maximum* of a continuous extension of the submodular function, and then using the pipage rounding of Ageev and Sviridenko [AS04] to return a discrete solution which achieves the $(1-e^{-1})-$approximation factor. For our problem, the greedy algorithm of Section 7.2 is preferable to obtain a $(1-e^{-1})$-approximation factor, but the ideas of Calinescu and his coauthors are useful to establish the approximability factor of the rank optimization problem (P_0) by rounding algorithms.

We also want to underline that the greedy algorithm may rise some computational issue when the number of parameters to estimate m is large. Fedorov [Fed72] suggested to make use of the Sherman-Morrison formula to speed up the computation. For $p = -1$ (resp. $p = 0$) indeed, i.e. for the A- (resp. D-) optimal design problem, one has to compute $\Phi_p(\mathcal{G}_k \cup \{i\})$ for each experiment i which is not yet in $\mathcal{G}_k$ at the k^{th} stage of the greedy process. This requires the computation of the inverse (resp. the determinant) of a $m \times m-$matrix, which is a very time-consuming task. Instead, the Sherman-Morrison formula allows one to compute the value of the increment thanks to a small-rank update. However, when working with arbitrary values of p, we cannot use these smart updates anymore. So at

the k^{th} stage of the greedy algorithm, one has to compute the m eigenvalues of $(s-k)$ information matrices, which is not practicable when m (the dimension of the parameter $\boldsymbol{\theta}$) is large (typically larger than 10000 in network applications).

In the sequel, we focus on the case in which $p=0$, and we consider the unit-cost case, where the number of experiments to select is n. We further assume without loss of generality that there is no prior measurement on the parameter ($A_0 = 0$). Note that we may always reduce to this case by defining the augmented observation matrices $\tilde{A}_i := [A_0^T/\sqrt{n}, A_i^T]^T$, so that we have

$$\sum_{i=1}^{s} w_i \tilde{A_i}^T \tilde{A}_i = A_0^T A_0 + \sum_{i=1}^{s} w_i A_i^T A_i.$$

7.3.1 A continuous relaxation

The continuous relaxation of the $D-$optimal problem is obtained by removing the integer constraint $\boldsymbol{w} \in \{0,1\}^s$:

$$\max_{\substack{\boldsymbol{w} \geq \mathbf{0} \\ \sum_k w_k \leq n}} \quad \det\Big(\sum_k w_k A_k^T A_k\Big). \tag{7.8}$$

We assume without loss of generality that the matrix $M(\mathbf{1}) = \sum_{k=1}^{s} A_k^T A_k$ is of full rank (where $\mathbf{1}$ denotes the vector of all ones), such that the optimal value of Problem (7.8) is positive. If this is not the case ($r^* := \operatorname{rank}(M(\mathbf{1})) < m$), we define instead a projected version of Problem (7.8): Let $U\Sigma U^T$ be a singular value decomposition of $M(\mathbf{1})$. We denote by U_{r^*} the matrix formed with the r^* leading singular vectors of $M(\mathbf{1})$, i.e. the r^* first columns of U. The $D-$optimal design problem is projected onto the observable space by mean of the projected observation matrices $\bar{A}_k := A_k U_{r^*}$ (see Paragraph 7.3 in [Puk93]):

$$\max_{\substack{\boldsymbol{w} \geq \mathbf{0} \\ \sum_k w_k \leq n}} \quad \det\Big(\sum_{k=1}^{s} w_k \bar{A_k}^T \bar{A}_k\Big). \tag{7.8'}$$

The function $\log(\det(\cdot))$ is strictly concave on the interior of $\mathbb{S}_m^+$, and Problem (7.8) can be solved by interior point techniques or multiplicative algorithms [Atw73, DPZ08, Yu10a, Sag09b]. The strict concavity of the logdet function indicates in addition that Problem (7.8) admits a unique solution if and only if

$$w_1 M_1 + w_2 M_2 + \ldots + w_s M_s = y_1 M_1 + y_2 M_2 + \ldots + y_s M_s \Rightarrow (w_1, \ldots, w_s) = (y_1, \ldots, y_s),$$

that is to say whenever the matrices $M_i = A_i^T A_i$ are linearly independent. In this chapter, we focus on the rounding techniques only, and we assume that an optimal solution $\boldsymbol{w}^*$ of the D-optimal design problem (7.8) is readily known. In the sequel, we also denote a discrete solution of Problem (P_0) by S^*. Since $M(\boldsymbol{w}^*)$ is of maximal rank r^*, we have:

$$r^* := \operatorname{rank}(M(\mathbf{1})) = \operatorname{rank}(M(\boldsymbol{w}^*)) = \varphi_0(\boldsymbol{w}^*) \geq \varphi_0(S^*). \tag{7.9}$$

The aim of this section is to propose some randomized rounding techniques which ascertain some approximation bounds. We clarify this statement in the following definition:

Definition 7.3.1. We say that an algorithm approximates the optimal solution of the rank optimization problem (P_0) by a factor α if for all possible instances, it returns a feasible random subset $\hat{S}$ such that:

$$\mathbb{E}(\varphi_0(\hat{S})) \geq \alpha\varphi_0(S^*).$$

Notice that, due to inequality (7.9), it is sufficient to show that $\mathbb{E}(\varphi_0(\hat{S})) \geq \alpha\varphi_0(\boldsymbol{w}^*) = \alpha r^*$ to prove that some rounding approximates the optimal solution by a factor α.

7.3.2 Roundings of the optimal solution

We now present two ingredients which will be useful in the sequel : the pipage rounding algorithm of Ageev and Sviridenko [AS04] and its relation with the extension by expectation of a submodular function, brought to light by Calinescu et. al. [CCPa07].

Extension by expectation and Pipage Rounding

We will make use of the extension by expectation [CCPa07] of a submodular set function φ, which is defined by

$$F_\varphi(\boldsymbol{y}) = \mathbb{E}[\varphi(\hat{S})], \tag{7.10}$$

where $\hat{S}$ is a random set of $[s]$ which contains $\{i\}$ independently with probability y_i. In other words,

$$F_\varphi(\boldsymbol{y}) = \sum_{S\subset\{1,..,s\}} \varphi(S) \prod_{i\in S} y_i \prod_{i\notin S}(1-y_i). \tag{7.11}$$

In our setting, we will denote by F_0 the extension by expectation of the rank function φ_0. Note that this extension can be defined only if all coordinates of $\boldsymbol{y}$ are smaller than 1. If $y_i > 1$ for some experiment i, we have to use another rounding technique, like the *proportional rounding* which we next present. Also note that if $\boldsymbol{y}$ is the $0/1$-vector associated to S, we have $F_\varphi(\boldsymbol{y}) = \varphi(S)$, which tells us that F_φ is an extension of φ indeed.

The idea of Calinescu et. al. (as reduced to the simple case of uniform matroids) is to find a vector $\boldsymbol{y}^*$ such that $F_\varphi(\boldsymbol{y}^*) \geq (1-1/e)\ OPT$, where OPT is the optimal value of the problem $\max_{|S|\leq n}\ \varphi(S)$. Then, they round $\boldsymbol{y}^*$ to a feasible discrete solution S with the *pipage rounding* algorithm of Ageev and Sviridenko [AS04], which satisfies with a high probability $\varphi(S) \geq F_\varphi(\boldsymbol{y}^*)$. Similarly, we will ask ourselves whether one can guarantee that $F_0(\boldsymbol{w}^*) \geq \alpha\varphi_0(S^*)$ for some α, in which case we could apply the pipage rounding technique to return a feasible subset S satisfying (with a high probability)

$$\varphi_0(S) = F_0(S) \geq F_0(\boldsymbol{w}^*) \geq \alpha\varphi_0(S^*).$$

For the reader's convenience, we now present the randomized version of the pipage rounding algorithm for the simple case of uniform matroids, and the ideas of the proof of Calinescu and his coauthors on the efficiency of this rounding ($\mathbb{E}[\varphi(S)] \geq F_\varphi(\boldsymbol{w}^*)$). Assume that we are given a nonnegative vector $\boldsymbol{y} \in [0,1]^s$ such that $\sum_i y_i = n$, and two indexes i and j for which y is fractional. The idea of this rounding technique is based on the fact that, for any submodular function φ, the function $F_{ij}^{\boldsymbol{y}} : t \mapsto F_\varphi\Big(\boldsymbol{y} + t(\mathbf{e}_i - \mathbf{e}_j)\Big)$ is convex [CCPa07], such that F_φ is increasing when we move in one of the directions $(\mathbf{e}_i - \mathbf{e}_j)$ or $(\mathbf{e}_j - \mathbf{e}_i)$. Therefore, we can increase one of the two variables (y_i or y_j) and decrease the other one until y_i or y_j becomes a 0 or a 1. Moreover, the sum of the vector is preserved along this transformation, which guarantees that the set obtained with this rounding will satisfy the desired property ($|S| = n$). In the randomized version (Algorithm 7.3.1), we choose between the two admissible directions with probabilities which ensure that we do not loose in expectation. This avoids costly evaluations of $F_\varphi(\boldsymbol{y})$.

Lemma 7.3.2 (Calinescu et al [CCPa07]). *Given a vector $\boldsymbol{y} \in [0,1]^s$ such that $\sum_i y_i = n$, **PipageRound($\boldsymbol{y}$)** returns in s iterations a random set S of cardinality n, of expected value $\mathbb{E}[\varphi(S)] \geq F_\varphi(\boldsymbol{y})$.*

Proportional Rounding

We now present another rounding scheme, which can be used even if some coordinates of $\boldsymbol{y}$ are larger than 1. The principle of this rounding is to start with $S_0 = \emptyset$, and, for $k = 1, ..., n$, we construct S_k from S_{k-1} by adding in it exactly one new element, namely $i \in [s] \setminus S_{k-1}$ with probability $\frac{y_i}{\sum_{j \notin S_{k-1}} y_j}$. If at some point, all the remaining coordinates $(y_j)_{j \notin S_{k-1}}$ are equal to 0, uniform probabilities are used. An alternative way to define this rounding is to generate a random vector X, the i^{th} coordinate of which is following an independent exponential distribution of expected value $1/y_i$: $X_i \sim \exp(1/y_i)$. As a consequence of the memoryless property of the exponential distribution, the set S_n can be

Algorithm 7.3.1 PipageRound($\boldsymbol{y}$)

Input: $\boldsymbol{y} \in [0,1]^s$ such that $\sum_i y_i = n$
while $\boldsymbol{y}$ is not integral **do**
  Pick i, j such that y_i and y_j are not in $\{0,1\}$.
  $\boldsymbol{\varepsilon} \leftarrow \{y_j, -y_i, 1 - y_i, y_j - 1\}$
  $\varepsilon^+ \leftarrow \min\{\varepsilon_i | \varepsilon_i > 0\}$
  $\varepsilon^- \leftarrow \max\{\varepsilon_i | \varepsilon_i < 0\}$
  $p \leftarrow \frac{\epsilon^+}{\epsilon^+ - \epsilon^-}$
  with probability p
    $y_i \leftarrow y_i + \epsilon^-, \quad y_j \leftarrow y_j - \epsilon^-$
  else
    $y_i \leftarrow y_i + \epsilon^+, \quad y_j \leftarrow y_j - \epsilon^+$
end while
Output: $\boldsymbol{y}$.

generated by selecting the indexes of the n smallest elements in the vector X (we use the convention $1/0 = \infty$, and if $\boldsymbol{y}$ has no more than n positive components we choose with uniform probabilities between the indices of X such that $X_i = \infty$).

We denote by $S_n(\boldsymbol{y})$ the random set of cardinality n obtained by this procedure, which we call *proportional rounding* of vector $\boldsymbol{y}$.

7.3.3 Characterization of $D-$optimality

We now give a characterization of the $D-$optimal design. This proposition is known as the *General Equivalence Theorem* in the full rank case, and was first stated by Fedorov [Fed72] for multiresponse experiments (cf. Chapter 2). We show here that it can also be stated in the degenerate case (where $\mathrm{rank}(M(\mathbf{1}) = r^* < m)$) with the help of generalized Moore-Penrose inverses.

Proposition 7.3.3 (General Equivalence Theorem)**.** *The design $\boldsymbol{w}^*$ is D-optimal (i.e. $\boldsymbol{w}^*$ is a solution of Problem* (7.8')*, which reduces to* (7.8) *in the full rank case $r^* = m$) if and only if for all $i \in [s]$, we have either:*

- $w_i^* = 0$
- *or* $w_i^* > 0$, *and* $\mathrm{trace}\ A_i M(\boldsymbol{w}^*)^\dagger A_i^T = \frac{\varphi_0(\boldsymbol{w}^*)}{n} = \frac{r^*}{n}$.

Proof. This proposition is known as the *General Equivalence Theorem* in the full rank case (where $r^* = m$, and the Moore-Penrose inverse is a regular inverse). For a proof, see Fedorov [Fed72], who deals with the normalized constraint ($n = 1$). The generalization to an arbitrary value of n is straightforward.

We now study the degenerate case, where $r^* < m$, and the $D-$optimal design is the solution of Problem (7.8'). The projected observation matrices $\bar{A}_k$ satisfy the full rank property by definition ($\overline{M(\mathbf{1})} := \sum_k \bar{A}_k{}^T \bar{A}_k$ is of size $r^* \times r^*$ and has rank r^*). This allows us to apply the full rank general equivalence theorem to characterize $\boldsymbol{w}^*$: the design $\boldsymbol{w}^*$ is $D-$optimal if and only if for all $i \in [s]$, we have either $w_i^* = 0$, or

$$\mathrm{trace}\ \bar{A}_i \overline{M(\boldsymbol{w}^*)}^{-1} \bar{A}_i{}^T = \frac{r^*}{n}, \tag{7.12}$$

where $\overline{M(\boldsymbol{w}^*)} := \sum_k w_k \bar{A}_k{}^T \bar{A}_k = U_{r^*}^T M(\boldsymbol{w}^*) U_{r^*}$. Since the range of $M(\boldsymbol{w}^*)$ is included in the one of $M(\mathbf{1})$, we have:

$$M(\boldsymbol{w}^*) = U \left(\begin{array}{c|c} \overline{M(\boldsymbol{w}^*)} & 0 \\ \hline 0 & 0 \end{array} \right) U^T,$$

where the diagonal blocks are of size $r^* \times r^*$ and $(m - r^*) \times (m - r^*)$ respectively. We can now express the Moore-Penrose inverse of $M(\boldsymbol{w}^*)$:

$$M(\boldsymbol{w}^*)^\dagger = U \left(\begin{array}{c|c} \overline{M(\boldsymbol{w}^*)}^{-1} & 0 \\ \hline 0 & 0 \end{array} \right) U^T = U_{r^*} \overline{M(\boldsymbol{w}^*)}^{-1} U_{r^*}^T.$$

Finally, we re-express the left hand side of (7.12), which will conclude the proof:

$$\operatorname{trace}\, \bar{A}_i \overline{M(\boldsymbol{w}^*)}^{-1} \bar{A}_i^{\,T} = \operatorname{trace}\, A_i U_{r^*} \overline{M(\boldsymbol{w}^*)}^{-1} U_{r^*}^T A_i^T = \operatorname{trace}\, A_i M(\boldsymbol{w}^*)^\dagger A_i^T.$$

□

We next give a proposition which shows how we can bound the components w_i^* of the $D-$optimal design. This was proved in a simpler case by Atwood [Atw73], who obtained $\frac{w_i^*}{n} \leq \frac{1}{m}$ when the observations are scalar (single response experiments), i.e. when the observation matrices are row vectors. The first part of the next result was discovered independently (in the regular case $r^* = m$) by Harman and Trnovská [HT09] (the latter article was published shortly after we had submitted an announcement of the present results to the conference ISCO 2010 [BGS10]). The proof of our result also adapts to the case in which the experimenter wants to estimate a subsystem $K^T\boldsymbol{\theta}$ of the parameters (cf. Theorem 2.4.7).

Proposition 7.3.4. *Let $\boldsymbol{w}^*$ be a $D-$optimal design. For all $i \in [s]$, we have the following bound on the optimal coordinate w_i^*:*

$$\frac{w_i^*}{n} \leq \frac{\operatorname{rank} M_i}{\operatorname{rank}(\sum_{i=1}^n M_i)}, \tag{7.13}$$

where $M_i := A_i^T A_i$. More generally, for an arbitrary subset S of $[s]$,

$$\frac{\sum_{i\in S} w_i^*}{n} \leq \frac{\operatorname{rank}(\sum_{i\in S} M_i)}{\operatorname{rank}(\sum_{i=1}^n M_i)} = \frac{\varphi_0(S)}{\varphi_0(\boldsymbol{w}^*)}. \tag{7.14}$$

Proof. The first inequality is trivial when $w_i^* = 0$. For any other value of $w_i^* > 0$, we make use of the characterization of optimality from the general equivalence theorem:

$$\operatorname{trace}\, A_i M(\boldsymbol{w}^*)^\dagger A_i^T = \frac{r^*}{n}.$$

Now, we replace $M(\boldsymbol{w}^*)^\dagger$ by $M(\boldsymbol{w}^*)^\dagger M(\boldsymbol{w}^*) M(\boldsymbol{w}^*)^\dagger$ in the right hand side of this expression, and we obtain:

$$\begin{aligned}
\frac{r^*}{n} &= \operatorname{trace} A_i M(\boldsymbol{w}^*)^\dagger \Big(\sum_k w_k^* A_k^T A_k \Big) M(\boldsymbol{w}^*)^\dagger A_i^T \\
&= \sum_k w_k^* \operatorname{trace} \underbrace{A_i M(\boldsymbol{w}^*)^\dagger A_k^T}_{X(i,k)} A_k M(\boldsymbol{w}^*)^\dagger A_i^T \\
&= \sum_k w_k^* \operatorname{trace} X(i,k) X(i,k)^T \\
&\geq w_i^* \operatorname{trace} X(i,i) X(i,i)^T,
\end{aligned} \tag{7.15}$$

where the inequality follows from the fact that the trace of any semidefinite matrix is nonnegative.

Let r_i denote the the rank of M_i, such that there exists a $r_i \times m$ matrix H_i such that $M_i = H_i^T H_i$. We have:

$$\begin{aligned}\operatorname{trace} X(i,i)X(i,i)^T &= \operatorname{trace}\ A_i M(\boldsymbol{w}^*)^\dagger A_i^T A_i M(\boldsymbol{w}^*)^\dagger A_i^T \\ &= \operatorname{trace}\ \underbrace{H_i M(\boldsymbol{w}^*)^\dagger H_i^T}_{\tilde{X}_i} H_i M(\boldsymbol{w}^*)^\dagger H_i^T.\end{aligned}$$

Now, notice that $\tilde{X}_i$ is a $r_i \times r_i$ symmetric matrix which has trace r^*. This allows us to write:

$$\operatorname{trace}(\tilde{X}_i \tilde{X}_i^{\ T}) = \sum_{j,k} \tilde{X}^2_{i_{(j,k)}} \geq \sum_{j=1}^{r_i} \tilde{X}^2_{i_{(j,j)}}.$$

This latter expression is the sum of squares of elements which sum to r^*, and is minimized when all these elements are equal, i.e. whenever $\tilde{X}_{i_{(j,j)}} = r^*/r_i$. Finally,

$$\operatorname{trace}(\tilde{X}_i \tilde{X}_i^{\ T}) \geq \sum_{j=1}^{r_i} \left(\frac{r^*}{r_i}\right)^2 = \frac{r^{*2}}{r_i}.$$

Inserting this lower bound in (7.15), we finally obtain $r^* \geq w_i^* \frac{r^{*2}}{r_i}$, or equivalently

$$w_i^* \leq \frac{r_i}{r^*} = \frac{r_i}{\varphi_0(\boldsymbol{w}^*)},$$

and the first inequality is proved. In order to generalize this result, let S be a subset of $[s]$. We exclude the trivial case $\sum_{i\in S} w_i^* = 0$, and we define $M_S = \sum_{i\in S} \frac{w_i^*}{\sum_{j\in S} w_j^*} M_i$. We consider the problem

$$\begin{aligned}\max_{v_S,(v_k)_{k\notin S}} \quad & \det\left(v_S M_S + \sum_{k\notin S} v_k M_k\right) \\ s.t. \quad & v_S + \sum_{k\notin S} v_k \leq n \\ & v_S \geq 0, \quad \forall k \notin S,\ v_k \geq 0.\end{aligned} \tag{7.16}$$

(Eventually, we may instead consider the projected problem with $\bar{M}_i = U_{r^*}^T M_i U_{r^*}$ if we are in the degenerate case $r^* < m$). It is clear that this problem has solution $v_k^* = w_k^*$ for $k \notin S$, and $v_S^* = \sum_{j\in S} w_j^*$, since any better value would contradict the $D-$optimality of $\boldsymbol{w}^*$. Applying the first inequality, we find:

$$\frac{\sum_{j\in S} w_j^*}{n} = \frac{v_S^*}{n} \leq \frac{\operatorname{rank}\ M_S}{r^*} \leq \frac{\operatorname{rank}\ (\sum_{i\in S} M_i)}{r^*}.$$

□

7.3.4 Rounding approximation factor for rank-optimality

Before we give the approximation factor that one can guarantee by using our rounding procedures, we need the two following technical lemmas.

Lemma 7.3.5. *Let $\alpha\Delta_s$ denote the simplex $\{\boldsymbol{x} \in (\mathbb{R}_+)^s \,|\, \sum_i x_i = \alpha\}$. We define the random variable $W_n(\boldsymbol{w}) = \sum_{i \in S_n(\boldsymbol{w})} w_i$, where $S_n(\boldsymbol{w})$ is the random subset of $[s]$ obtained by proportional rounding. Then, we have*

$$\forall w \in \alpha\Delta_s, \quad \mathbb{E}[W_n(\boldsymbol{w})] \geq \mathbb{E}[W_n(\frac{\alpha}{s}, ..., \frac{\alpha}{s})] = n\frac{\alpha}{s}.$$

Proof. First notice that we can give the expression of $\mathbb{E}[W_n(\boldsymbol{w})]$ in close form by summing over all permutation of n elements in $[s]$:

$$\mathbb{E}[W_n(\boldsymbol{w})] = \sum_{\sigma \in \pi(n,s)} \frac{w_{\sigma_1}}{\sum_i w_i} \cdot \frac{w_{\sigma_2}}{\sum_{i \neq \sigma_1} w_i} \cdots \frac{w_{\sigma_n}}{\sum_{i \notin \{\sigma_1, \ldots, \sigma_{n-1}\}} w_i} \cdot (w_{\sigma_1} + \ldots + w_{\sigma_n}).$$

Although this expression looks particularly awful, the reader can verify that it can be obtained by the following induction procedure:

$$\begin{cases} \mathbb{E}[W_1(\boldsymbol{w})] & = \frac{\sum_i w_i^2}{\sum_i w_i} \\ \mathbb{E}[W_{k+1}(\boldsymbol{w})] & = \frac{1}{\sum_j w_j} \sum_{i=1}^{s} w_i(w_i + \mathbb{E}[W_k(\boldsymbol{w}_{\setminus\{i\}})]) \end{cases},$$

where $\boldsymbol{w}_{\setminus\{i\}}$ is the vector of length $s-1$ with entries $(w_1, \ldots, w_{i-1}, w_{i+1}, \ldots, w_s)$. The latter formula is easily obtained by considering the expansion of a probability tree, and will allow us to make a proof by induction. We are going to show that $\forall k \leq s$, $\mathbb{E}[W_k(\boldsymbol{w})]$ attains its minimum value $k\frac{\alpha}{s}$ on $\alpha\Delta_s$ for the uniform vector. For $k = 1$, $\mathbb{E}[W_1(\boldsymbol{w})] = \frac{1}{\alpha}\sum_i w_i^2$ on the $\alpha-$ simplex, which is a convex and symmetric function, the minimum of which is attained for the uniform vector:

$$\mathbb{E}[W_1(\frac{\alpha}{s}, \ldots, \frac{\alpha}{s})] = \frac{s\left(\frac{\alpha}{s}\right)^2}{s\left(\frac{\alpha}{s}\right)} = \frac{\alpha}{s}.$$

Now, we assume that the statement is true for a given $k \in \{1, \ldots, s-1\}$:

$$\forall \boldsymbol{w} \in \alpha\Delta_s, \quad , \mathbb{E}[W_k(\boldsymbol{w})] \geq \mathbb{E}[W_k(\frac{\alpha}{s}, \ldots, \frac{\alpha}{s})] = k\frac{\alpha}{s}.$$

Let $\boldsymbol{w} \in \alpha\Delta_s$. For all $i \in [s]$, the vector $\boldsymbol{w}_{\setminus\{i\}}$ is in the simplex $(\alpha - w_i)\Delta_{s-1}$. So, using our induction hypothesis, we find :

$$\mathbb{E}[W_k(\boldsymbol{w}_{\setminus\{i\}})] \geq k\frac{\alpha - w_i}{s-1},$$

and using the inductive construction of $\mathbb{E}[W_{k+1}(\boldsymbol{w})]$,

$$\mathbb{E}[W_{k+1}(\boldsymbol{w})] \geq \underbrace{\frac{1}{\alpha}\sum_{i=1}^{s} w_i(w_i + k\frac{\alpha - w_i}{s-1})}_{g_k(\boldsymbol{w})}.$$

It is clear that g_k is symmetric. Moreover, we can see that g_k is convex since it is a separable function and for $k < s$,

$$\frac{\partial^2 g(\boldsymbol{w})}{\partial w_i^2} = \frac{2}{\alpha}(1 - \frac{k}{s-1}) \geq 0.$$

This shows that, for $k < s$ the minimum of g_k is attained on the $\alpha-$simplex for the uniform vector $(\frac{\alpha}{s}, \ldots, \frac{\alpha}{s})$. This gives the following lower bound on $\mathbb{E}[W_{k+1}(\boldsymbol{w})]$:

$$\forall w \in \alpha\Delta_s, \quad \mathbb{E}[W_{k+1}(\boldsymbol{w})] \geq g(\frac{\alpha}{s}, \ldots, \frac{\alpha}{s}) = (k+1)\frac{\alpha}{s}.$$

Moreover, this bound is attained for the uniform vector, since it leads to consider an expected value on a uniform probability tree with $(k+1)\frac{\alpha}{s}$ on each extremal leaf.
By induction, we conclude that our induction hypothesis holds for all $k \leq s$, and in particular for $k = n$. □

Lemma 7.3.6. *For all vector $\boldsymbol{w} \in [0,1]^s$, the following equality holds:*

$$\sum_{S \subset \{1,..,s\}} \left(\sum_{i \in S} w_i\right) \prod_{i \in S} w_i \prod_{i \notin S} (1 - w_i) = \sum_{i=1}^{s} w_i^2$$

Proof. We proceed by induction on s: for $s = 1$, the equality is trivial, since the summation reduces to $S = \emptyset$ and $S = \{1\}$, and has only one nonzero term: w_1^2.
Now, we assume that the equality from this lemma is true for a given s, and we write (by separating between the sets which contains $\{s+1\}$ and those which do not).

$$\begin{aligned}
&\sum_{S \subset \{1,..,s+1\}} \left(\sum_{i \in S} w_i\right) \prod_{i \in S} w_i \prod_{i \notin S} (1 - w_i) \\
&= w_{s+1}\left(w_{s+1} + \sum_{S \subset \{1,..,s\}} \left(\sum_{i \in S} w_i\right) \prod_{i \in S} w_i \prod_{i \notin S} (1 - w_i)\right) \\
&\quad + (1 - w_{s+1})\left(\sum_{S \subset \{1,..,s\}} \left(\sum_{i \in S} w_i\right) \prod_{i \in S} w_i \prod_{i \notin S} (1 - w_i)\right) \\
&= \sum_{i=1}^{s} w_i^2 \left((1 - w_{s+1}) + w_{s+1}\right) + w_{s+1}^2 \\
&= \sum_{i=1}^{s+1} w_i^2,
\end{aligned}$$

where the induction hypothesis has been used to replace the summation over $S \subset [s]$ by $\sum_{i=1}^{s} w_i^2$. □

We can now formulate the main result of this section:

Theorem 7.3.7 (Rounding Approximability Factor)**.** *Let $\boldsymbol{w}^*$ be a $D-$optimal design. The proportional rounding of the vector $\boldsymbol{w}^*$ approximates the optimal solution of the rank maximization problem (P_0) by $\frac{n}{s}$. Moreover, if all coordinates of $\boldsymbol{w}^*$ are smaller than 1, then the pipage rounding algorithm gives the same approximation factor of $\frac{n}{s}$.*

Proof. We first point out that if $\boldsymbol{w}^*$ has no more than n positive entries, $S_n(\boldsymbol{w}^*)$ always contains the indices of these entries, such that the rounded design $S_n(\boldsymbol{w}^*)$ is of maximal rank: $\varphi_0(S_n(\boldsymbol{w}^*)) = r^*$, and the approximation ratio is 1. Otherwise, we bound the approximation ratio $\frac{\mathbb{E}[\varphi_0(S_n(\boldsymbol{w}^*))]}{\varphi_0(\boldsymbol{w}^*)}$ thanks to the result of Proposition 7.3.4 :

$$\begin{aligned}\frac{\mathbb{E}[\varphi_0(S_n(\boldsymbol{w}^*))]}{\varphi_0(\boldsymbol{w}^*)} &= \sum_{\sigma\in\pi(n,s)} \frac{w^*_{\sigma_1}}{\sum_i w^*_i}\cdot\frac{w^*_{\sigma_2}}{\sum_{i\neq\sigma_1} w^*_i}\cdots\frac{w^*_{\sigma_n}}{\sum_{i\notin\{\sigma_1,\ldots,\sigma_{n-1}\}} w^*_i}\cdot\frac{\varphi_0(\sigma)}{\varphi_0(\boldsymbol{w}^*)}.\\ &\geq \sum_{\sigma\in\pi(n,s)} \frac{w^*_{\sigma_1}}{\sum_i w^*_i}\cdot\frac{w^*_{\sigma_2}}{\sum_{i\neq\sigma_1} w^*_i}\cdots\frac{w^*_{\sigma_n}}{\sum_{i\notin\{\sigma_1,\ldots,\sigma_{n-1}\}} w^*_i}\cdot\frac{(w^*_{\sigma_1}+\ldots+w^*_{\sigma_n})}{n}\\ &= \frac{1}{n}\mathbb{E}[W_n(\boldsymbol{w}^*)].\end{aligned}$$

In the above, the summation is taken over the $\frac{s!}{(s-n)!}$ permutations σ of n elements in $[s]$, and $W_n(\boldsymbol{w}^*)$ is the random variable which has been defined in Lemma 7.3.5. Since $\boldsymbol{w}^*$ is in the $n-$simplex, we obtain the desired approximation factor from Lemma 7.3.5:

$$\frac{\mathbb{E}[\varphi_0(S_n(\boldsymbol{w}^*))]}{\varphi_0(\boldsymbol{w}^*)} \geq \frac{1}{n}\frac{n^2}{s} = \frac{n}{s}.$$

Similarly, if all coordinates of $\boldsymbol{w}^*$ are smaller than 1, then the extension by expectation F_0 is well defined at $\boldsymbol{w}^*$, and by Lemma 7.3.6:

$$\begin{aligned}\frac{F_0(\boldsymbol{w}^*)}{\varphi_0(\boldsymbol{w}^*)} &= \sum_{S\subset\{1,..,s\}} \frac{\varphi_0(S)}{\varphi_0(\boldsymbol{w}^*)}\prod_{i\in S} w^*_i\prod_{i\notin S}(1-w^*_i)\\ &\geq \sum_{S\subset\{1,..,s\}} \frac{\sum_{i\in S} w^*_i}{n}\prod_{i\in S} w^*_i\prod_{i\notin S}(1-w^*_i)\\ &= \frac{1}{n}\sum_{i=1}^{s} w_i^{*2}\\ &\geq \frac{s}{n}\Big(\frac{n}{s}\Big)^2 = \frac{n}{s},\end{aligned}$$

where the latter inequality is once again the minimality of $x\mapsto\sum_{i=1}^s x_i^2$ over $n\Delta_s$ for $w=(\frac{n}{s},...,\frac{n}{s})$. Hence, the pipage rounding approximates the optimal solution within a factor of $\frac{n}{s}$, thanks to Lemma 7.3.2. □

Remark 7.3.1. The inequalities $\mathbb{E}[\varphi_0(S_n(\boldsymbol{w}^*))]\geq\frac{n}{s}\varphi_0(\boldsymbol{w}^*)$ and $F_0(\boldsymbol{w}^*)\geq\frac{n}{s}\varphi_0(\boldsymbol{w}^*)$ are optimal. The reader can verify indeed that they are attained for the following $s\times s-$

observation matrices:

$$M_1 = \begin{pmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix}, \quad M_2 = \begin{pmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix}, \ldots, \quad M_s = \begin{pmatrix} 0 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix}.$$

In the last theorem we give an approximation factor by comparing the expected value of φ_0 for the rounded set to $\varphi_0(\boldsymbol{w}^*)$. The reader may ask himself if these bounds are accurate, since the approximation factor of a rounding algorithm is actually defined with respect to the discrete optimal value $\varphi_0(S^*)$. We answer partially with these two remarks:

Remark 7.3.2. For $s > n+1$, we can find observation matrices for which the ratios $\frac{\mathbb{E}[\varphi_0(S_n(\boldsymbol{w}^*))]}{\varphi_0(S^*)}$ and $\frac{F(\boldsymbol{w}^*)}{\varphi_0(S^*)}$ take the value $\frac{n}{s-1}$. This indicates that the optimal approximation factor is somewhere between $\frac{n}{s}$ and $\frac{n}{s-1}$. Consider the following $(s-1)\times(s-1)-$ observation matrices indeed:

$$M_1 = \left(\begin{array}{c|c} 0 & 0 \\ \hline 0 & \varepsilon \boldsymbol{I} \end{array}\right), \; M_2 = \begin{pmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix}, \; M_3 = \begin{pmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix}, \ldots, \; M_s = \begin{pmatrix} 0 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix},$$

where the nonzero block in M_1 is of size $(s-n) \times (s-n)$. The reader can easily verify that for $\varepsilon < \frac{1}{s-n}$, $w_1^* = 0$, and $w_2^* = \ldots = w_s^* = \frac{n}{s-1}$, while the discrete solution S^* of Problem (7.8) is clearly $\{1, \ldots, n\}$, which is the only subset of n matrices that sums to a full rank matrix. Hence, this example yields an approximation factor of $\frac{n}{\mathrm{rank}(M_1+\ldots+M_n)} = \frac{n}{s-1}$ for both the proportional and the pipage rounding.

Remark 7.3.3. for $n = 1$ and $s > 2$, we can show that $\frac{n}{s-1}$ is the optimal approximation factor for the proportional rounding algorithm. Since $n = 1$, the discrete optimum S^* of Problem (7.8) is a singleton, which we can consider to be $\{1\}$ without loss of generality. Now, we bound the the approximation ratio:

$$\begin{aligned} \frac{\mathbb{E}[\varphi_0(S_1(\boldsymbol{w}^*))]}{\varphi_0(\{1\})} &= \sum_{i=1}^{s} w_i^* \frac{\varphi_0(\{i\})}{\varphi_0(\{1\})} \\ &\geq w_1^* + \sum_{i=2}^{s} w_i^* \frac{\varphi_0(\{i\})}{\varphi_0(\boldsymbol{w}^*)} \\ &\geq w_1^* + \sum_{i=2}^{s} (w_i^*)^2, \end{aligned}$$

where the first inequality follows from $\varphi_0(\{1\}) \leq \varphi_0(\boldsymbol{w}^*)$, and the second one from Proposition 7.3.4. Now, using the fact that $\sum_{i=2}^{s}(w_i^*)^2$ is minimized on the $(1-w_1^*)-$simplex for the uniform vector $(w_2^* = \ldots = w_s^* = \frac{1-w_1^*}{s-1})$, we have:

$$\frac{\mathbb{E}[\varphi_0(S_1(\boldsymbol{w}^*))]}{\varphi_0(\{1\})} \geq w_1^* + \frac{(1-w_1^*)^2}{s-1}.$$

The left hand side of this equation is an increasing function of $\boldsymbol{w}^*$ on $[0,1]$, such that we obtain the lower bound for $w_1^* = 0$:

$$\frac{\mathbb{E}[\varphi_0(S_1(\boldsymbol{w}^*))]}{\varphi_0(\{1\})} \geq \frac{1}{s-1}.$$

In the above discussion, we characterized the rounding approximation factor for Problem (P_p) when $p \to 0$. Our proof does not seem to adapt for other values of $p \in]0,1]$, but we think that Proposition 7.3.4 might adapt to other values of p in the following way:
Let $p \in [0,1]$ and let $\boldsymbol{w}^*$ be optimal for the continuous relaxation of Problem (P_p). Is it true that for an arbitrary subset S of $[s]$,

$$\frac{\sum_{i \in S} w_i^{*(1-p)}}{n} \leq \frac{\varphi_p(S)}{\varphi_p(\boldsymbol{w}^*)} \ ?$$

We leave it here as an open question, but we underline that, following the same reasoning as above, this would provide an approximation factor of $\left(\frac{n}{s}\right)^{1-p}$ for Problem $(P_p),\ p \in [0,1]$. Interestingly, this bound is attained for diagonal observation matrices with disjoint support. Note that this formula would show that there is a continuously increasing difficulty from the easy case $(p = 1)$ to the most degenerate problem $(p = 0)$.

7.4 Conclusion

This chapter gives bounds on the behavior of some classical heuristics used for combinatorial problems arising in optimal experimental design. Our results can either justify or discard the use of such heuristics, depending on the settings of the instances considered. Moreover, our results confirm some facts that had been observed in the literature, namely that rounding algorithms perform better if the density of measurements is high, and that the greedy algorithm always gives a quite good solution. We illustrate these observations with two examples:

In a sensor location problem, Uciński and Patan [UP07] noticed that the trimming of a Branch and Bound algorithm was better if they activated more sensors, although this led to a much larger research space. The authors claims that this surprising result can be explained by the fact that a higher density of sensors leads to a better continuous relaxation. This is confirmed by our result of approximability, which shows that the larger is the number of selected experiments, the better is the quality of the rounding.

It is also known that the greedy algorithm generally gives very good results for the optimal design of experiments (see e.g. [SQZ06], where the authors explicitly chose not to implement a local search from the design greedily chosen, since the greedy algorithm already performs very well). Our $(1 - 1/e)-$approximability result guarantees that this algorithm always well behaves indeed.

Part II

Optimal monitoring in large Networks

Chapter 8

Inference of the traffic matrix: a review

The traffic matrix (TM) of a network gives the volume of traffic between all pairs of origin and destination nodes of a network. This matrix is a crucial input for many network planning operations, and its estimation is therefore an essential problem. For example, the routing table, which specifies the path between every pair of origin and destination, should clearly be decided with an accurate prevision of the demand in order to avoid congestion. Similarly, the traffic matrix is a deciding piece of information when an Internet Service Provider (ISP) decides to upgrade the capacity of a link on its network. Other important applications of the traffic matrix include anomaly detection, billing and development of failover strategies.

However, the inference of traffic matrices turns out to be a difficult problem. The estimation of traffic matrices in networks has therefore attracted much interest for the last decade, from both Internet providers and the network research community. In this chapter, we shall review the different methods that have been proposed for this task; they can principally be classified in two types: those relying on the link counts only, and those which take advantage of direct network measurements provided by a monitoring software. We also indicate the reviews of Benameur and Roberts [BR04], and Vaton, Bedo and Gravey [VBG05], which cover some of the techniques presented in this chapter.

8.1 Notation and definitions

We refer as *traffic matrix* the set of volumes of traffic on each Origin-Destination (OD) pair of a network, during a given time interval whose typical length varies from five minutes to one hour. On a network with n nodes (routers), this data can indeed be represented by a $n \times n$ matrix, the $(o, d)-$entry of which corresponds to the volume of traffic from Node o to Node d (during the given time interval). In the practice, we often rearrange this matrix as a vector $\boldsymbol{x}$ of length $m = n^2$ to facilitate the notation, but we still refer this vector as the *traffic matrix*, and we shall sometimes continue to use the double indexing notation $x_{o,d}$.

This vector notation also allows one to handle the case in which $m < n^2$ OD pairs are of interest (and we use a vector $\boldsymbol{x}$ of length m).

The traffic matrix is a dynamic object, since traffic volumes are evolving over time. When working over a global observation period which is divided in T time intervals, the unknown thus consists in the $m \times T-$matrix X, the columns of which are the vectors $\boldsymbol{x_1}, \ldots, \boldsymbol{x_T}$, where $\boldsymbol{x_t}$ represents the traffic matrix during the t^{th} time interval ($\boldsymbol{x_t}$ is a *snapshot* of the traffic matrix at time t). We shall still refer to X as the *traffic matrix*, or sometimes as the *dynamic traffic matrix*. The elements of $\boldsymbol{x_t}$ are denoted by $x_{o,d}^{(t)}$ and will be referred as the *flow volumes* (at time t) – these, however, should not be confused with the classic $5-$tupple flows from the networking literature, which refer to packets sharing the same source address, destination address, source port, destination port, and IP protocol.

8.2 Traffic matrix estimation from link counts

In the classic problem, we consider a network with n nodes and l links. Link measurements are provided by the Simple Network Management Protocol (SNMP), which gives some statistics on the links (for instance, the number of bytes seen on each link in a time window). An analogy with road traffic can be useful: in this case the link counts correspond to the number of vehicles seen on each road segment (during a time interval), and can be gathered thanks to pneumatic tubes or magnetic loops. We will denote the vector of SNMP link counts by $\boldsymbol{y}^{\mathbf{SNMP}} = (y_1, \ldots, y_l)^T$. Again, when the observation period is divided in T time intervals, we concatenate the measurements into a $l \times T-$matrix: $Y^{\text{SNMP}} = [\boldsymbol{y_1^{\mathbf{SNMP}}}, \ldots, \boldsymbol{y_T^{\mathbf{SNMP}}}]$, where $\boldsymbol{y_t^{\mathbf{SNMP}}}$ is the vector of link counts at time t (i.e. during the t^{th} time interval).

We are also given the set of m OD pairs of interest (usually, $m = n^2$), and for each pair, the set of links that a byte need traverse to go from Origin o to Destination d. The information about the routing is assumed to be known, and is classically gathered in the $l \times m$ incidence matrix A: this is a $0/1-$matrix whose (i, r)-entry takes the value 1 if and only if the OD pair r traverses link i. More generally, the Internet provider routing policies may lead us to consider matrices in which $A_{i,r}$ is a real number representing the fraction of the traffic from OD pair r that traverses link i.

8.2.1 An ill-posed problem

The problem of estimating the traffic matrix $\boldsymbol{x}$ from the link counts $\boldsymbol{y}^{\mathbf{SNMP}}$ (or, in a dynamic framework, estimating X from Y^{SNMP}), has been studied since the late 1970's in the framework of road traffic (see e.g. Van Zuylan and Willumsen [ZW80]) or telephone networks (e.g. Krupp [Kru79]). This work was a valuable source of inspiration for the information theoretic approach which we present below.

If we assume that the measurements are perfect, the following relation is easily seen to hold:

$$\boldsymbol{y}^{\text{SNMP}} = A\boldsymbol{x}. \tag{8.1}$$

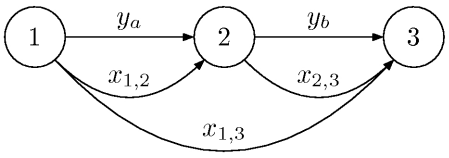
In typical networks, l is in the order of n, while m is in the order of n^2, such that the routing matrix A has more columns than rows, and the estimation of the traffic matrix $\boldsymbol{x}$ is an ill-posed problem (cf. Example 8.2.1). For the dynamic problem, the relation $Y^{\text{SNMP}} = AX$ is true if the routing matrix A remains the same during the whole observation period. If this is not the case, we have instead $\boldsymbol{y}_{\boldsymbol{t}}^{\text{SNMP}} = A_t\boldsymbol{x}_{\boldsymbol{t}}$ for all $t \in [T]$, where A_t is the routing matrix during the t^{th} time interval.

8.2.2 The information theoretic approach

After an appropriate normalization, the vector of OD flows $\boldsymbol{x}$ can be handled as a probability distribution defined on the OD pairs. This suggests to use the principle of minimum entropy to complete the partial information on $\boldsymbol{x}$ which is given by Equation (8.1). This approach is detailed in Chapter 9: in absence of any other information, the *traffic matrix of minimal entropy* which respects the ingress/egress measurements is known as the gravity model $\boldsymbol{x}^{\boldsymbol{G}}$, in which the traffic from o to d is proportional to the product of the incoming

---

Example 8.2.1. Here is a small toy example, to illustrate how we obtain the measurement equations:



The incidence table between the ODs and the links of this network is:

	OD $1 \to 2$	OD $2 \to 3$	OD $1 \to 3$
link a	1	0	1
link b	0	1	1

and one can easily verify that the vector of link counts $\boldsymbol{y} = [y_a, y_b]^T$ must satisfy

$$\boldsymbol{y} = \underbrace{\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}}_{A} \boldsymbol{x}, \qquad \text{where} \quad \boldsymbol{x} = \begin{bmatrix} x_{1,2} \\ x_{2,3} \\ x_{1,3} \end{bmatrix}.$$

In absence of any additional information on the vector of OD flows, we can only say that $\boldsymbol{x}$ belongs to the space of the nonnegative solutions of the latter equation:

$$\boldsymbol{x} = \begin{bmatrix} y_a - u \\ y_b - u \\ u \end{bmatrix} \quad \text{for a scalar } u \in [0, \min(y_a, y_b)].$$

traffic in o and the outgoing traffic at d:

$$x^G_{o,d} = \frac{x^{In}(o)x^{Out}(d)}{\sum_{i=1}^n x^{In}(i)},$$

where $x^{In}(i)$ (resp. $x^{Out}(i)$) denotes the total traffic entering the network (resp. exiting the network) at node i. Thinking about $x_{o,d}$ as the joint probability that a packet has the origin o and the destination d, it means that the source of a packet and its destination are independent. In practice, this model happens to be a good prior estimate for the real traffic matrix $\boldsymbol{x}$.

Zhang, Roughan, Lund and Donoho [ZRLD05] further proposed an extension of the gravity model, in which the ingress and egress links are separated in two classes: the class $\mathcal{C}$ of links serving *customers*, and the class $\mathcal{P}$ of those linked to *peers*. If we know for each ingress/egress link to which class it belongs, Zhang and his coauthors proposed a model in which the source and the destination of a packet are independent, *conditionally to the class of the source and the class of the destination*. Using the fact that there is no traffic transiting the network from one peer to another, they obtained the generalized gravity prior $\boldsymbol{x^{GG}}$:

$$x^{GG}_{o,d} = \begin{cases} 0 & \text{if} \quad o \in \mathcal{P},\ d \in \mathcal{P}; \\ x^{In}(o)x^{Out}(d)\frac{1}{\sum_{c\in\mathcal{C}} x^{Out}(c)} & \text{if} \quad o \in \mathcal{P},\ d \in \mathcal{C}; \\ x^{In}(o)x^{Out}(d)\frac{1}{\sum_{c\in\mathcal{C}} x^{In}(c)} & \text{if} \quad o \in \mathcal{C},\ d \in \mathcal{P}; \\ x^{In}(o)x^{Out}(d)\frac{\sum_{c\in\mathcal{C}} x^{In}(c) - \sum_{p\in\mathcal{P}} x^{Out}(p)}{\sum_{c\in\mathcal{C}x^{In}(c)} \sum_{c\in\mathcal{C}x^{Out}(c)}} & \text{if} \quad o \in \mathcal{C},\ d \in \mathcal{C}. \end{cases} \tag{8.2}$$

In a dynamic context, if we assume that the time intervals are short enough so that no big change occurs between two successive time steps, a natural prior for $\boldsymbol{x_t}$ is given by the estimation of the traffic at time $t-1$. This prior can then be projected (in the sense of entropy) on the feasible subspace $\boldsymbol{y_t^{\mathrm{SNMP}}} = A\boldsymbol{x_t}$, see Chapter 9. The resulting estimate is usually referred as the *tomogravity* estimate of the traffic matrix. We summarize this scheme of estimation of the traffic matrix in Algorithm 8.2.1, in which a parameter α is used to make a convex combination of the gravity prior and the previous estimate.

Algorithm 8.2.1 Dynamic estimation of the traffic matrix via entropic projections

Input: parameter $\alpha \in [0,1]$
for $t = 1, \ldots, T$ **do**
  Build the gravity estimate $\boldsymbol{x^G}$ (or generalized gravity $\boldsymbol{x^{GG}}$), with the SNMP data of time t;
  if t=1 **then**
    $\boldsymbol{x^{\mathrm{prior}}} \leftarrow \boldsymbol{x^G}$ (or $\boldsymbol{x^{GG}}$);
  else
    $\boldsymbol{x^{\mathrm{prior}}} \leftarrow \alpha\hat{\boldsymbol{x}}_{\boldsymbol{t-1}} + (1-\alpha)\boldsymbol{x^G}$ (or $\alpha\hat{\boldsymbol{x}}_{\boldsymbol{t-1}} + (1-\alpha)\boldsymbol{x^{GG}}$);
  end if
  Compute the estimation of the traffic $\hat{\boldsymbol{x}}_{\boldsymbol{t}}$ by projecting $\boldsymbol{x^{\mathrm{prior}}}$ onto the space $\{\boldsymbol{x} :\ \boldsymbol{y_t^{\mathrm{SNMP}}} = A\boldsymbol{x}\}$ (in the sense of entropy, see Chapter 9 for algorithms).
end for

8.2.3 The Bayesian approach

In the Bayesian approach, a simple parametric model for the flows is assumed, and we search the parameters which maximize the likelihood of the observations. Two class of models have been proposed: the Poisson model of Vardi [Var96], and the Gaussian model with a mean-variance relation of Cao et. al. [CDVY00]. Since both methods are similar, and Poisson distribution are approximated by Gaussian distribution in [Var96], we will only review the latter one.

Cao and his coauthors proposed a moving iid model on a sliding window of width h: for the estimation at time t, we assume that the vectors $\boldsymbol{x_{t-h}}, \ldots, \boldsymbol{x_t}, \ldots, \boldsymbol{x_{t+h}}$ are independent and identically distributed (iid) with a normal distribution $\mathcal{N}(\boldsymbol{\lambda_t}, \phi_t \operatorname{Diag}(\boldsymbol{\lambda_t})^c)$, where the exponent c is supposed to be known (the authors of [CDVY00] claim that a typical value for c is 2). Under these assumptions, the observations $\boldsymbol{y_{t-h}}, \ldots, \boldsymbol{y_t}, \ldots, \boldsymbol{y_{t+h}}$ are iid with distribution $\mathcal{N}(A\boldsymbol{\lambda_t}, A\Sigma_t A^T)$, where $\Sigma_t := \phi_t \operatorname{Diag}(\boldsymbol{\lambda_t})^c$, and the log-likelihood of these measurements is:

$$\begin{aligned}\ell\Big((\phi_t, \boldsymbol{\lambda_t})|\boldsymbol{y}\Big) = &-\frac{2h+1}{2}\log\det(A\Sigma_t A^T) \\ &-\frac{1}{2}\sum_{\tau=t-h}^{t+h}(\boldsymbol{y_\tau} - A\boldsymbol{\lambda_t})^T(A\Sigma_t A^T)^{-1}(\boldsymbol{y_\tau} - A\boldsymbol{\lambda_t}).\end{aligned}$$

The maximization of the latter expression with respect to $\boldsymbol{\lambda_t}$ and ϕ_t has no analytic solution and is a complicated problem. Instead, Cao et. al. [CDVY00] suggested to use the Expectation-Maximization (EM) algorithm [DLR77], for which convergence results toward a local maximum are known [Wu83]. The principle of this algorithm is to iteratively conduce an Expectation (E) step, in which the expectation of the log-likelihood $\ell\Big((\phi_t, \boldsymbol{\lambda_t})|\boldsymbol{x}\Big)$ is computed, conditionally to the observations $\boldsymbol{y_{t-h}}, \ldots, \boldsymbol{y_{t+h}}$ and the current estimate of the parameters $(\phi_t^{(k)}, \boldsymbol{\lambda_t^{(k)}})$:

$$Q\Big((\phi_t, \boldsymbol{\lambda_t})|(\phi_t^{(k)}, \boldsymbol{\lambda_t^{(k)}})\Big) = \mathbb{E}_{\boldsymbol{x}}\left[\ell\Big((\phi_t, \boldsymbol{\lambda_t})|\boldsymbol{x}\Big)|\boldsymbol{y}, \phi_t^{(k)}, \boldsymbol{\lambda_t^{(k)}}\right].$$

Then, a Maximization (M) step is applied in order to update the value of the current parameter:

$$(\phi_t^{(k+1)}, \boldsymbol{\lambda_t^{(k+1)}}) \leftarrow \operatorname*{argmax}_{\phi_t, \boldsymbol{\lambda_t}} Q\Big((\phi_t, \boldsymbol{\lambda_t})|(\phi_t^{(k)}, \boldsymbol{\lambda_t^{(k)}})\Big).$$

In fact, Cao et. al. showed that the E-step is analytic. The log-likelihood with respect to $\boldsymbol{x}$ takes indeed the form

$$\ell\Big((\phi_t, \boldsymbol{\lambda_t})|\boldsymbol{x}\Big) = -\frac{2h+1}{2}\log\det(\Sigma_t) - \frac{1}{2}\sum_{\tau=t-h}^{t+h}(\boldsymbol{x_\tau} - \boldsymbol{\lambda_t})^T\Sigma_t^{-1}(\boldsymbol{x_\tau} - \boldsymbol{\lambda_t}),$$

and for all $\tau \in \{t-h, \ldots, t+h\}$, the conditional distribution of $\boldsymbol{x_\tau}$ with respect to the observation $\boldsymbol{y_\tau}$ and the current estimate of the parameters $(\phi_t^{(k)}, \boldsymbol{\lambda_t^{(k)}})$ is Gaussian, with

mean and variance

$$\begin{aligned} \boldsymbol{m}_{t,\tau}^{(k)} &= \boldsymbol{\lambda}_t^{(k)} + \Sigma_t^{(k)} A^T (A\Sigma_t^{(k)} A^T)^{-1} (\boldsymbol{y}_\tau - A\boldsymbol{\lambda}_t^{(k)}); \\ R_t^{(k)} &= \Sigma_t^{(k)} - \Sigma_t^{(k)} A^T (A\Sigma_t^{(k)} A^T)^{-1} A\Sigma_t^{(k)}. \end{aligned}$$

Hence, we can give the function Q in close form:

$$\begin{aligned} Q\Big((\phi_t, \boldsymbol{\lambda_t})|(\phi_t^{(k)}, \boldsymbol{\lambda_t^{(k)}})\Big) &= -\frac{2h+1}{2}\log\det(\Sigma_t) \\ &\quad - \frac{1}{2}\sum_{\tau=t-h}^{t+h} \mathbb{E}_{\boldsymbol{x_\tau}}\left[(\boldsymbol{x_\tau} - \boldsymbol{\lambda_t})^T \Sigma_t^{-1} (\boldsymbol{x_\tau} - \boldsymbol{\lambda_t}) | \boldsymbol{y_\tau}, \phi_t^{(k)}, \boldsymbol{\lambda_t^{(k)}}\right] \\ &= -\frac{2h+1}{2}\log\det(\Sigma_t) \\ &\quad - \frac{1}{2}\sum_{\tau=t-h}^{t+h} \operatorname{trace}\Big(\Sigma_t^{-1} \underbrace{\mathbb{E}[\boldsymbol{x_\tau}\boldsymbol{x_\tau}^T | \boldsymbol{y_\tau}, \phi_t^{(k)}, \boldsymbol{\lambda_t^{(k)}}]}_{R_t^{(k)} + \boldsymbol{m}_{t,\tau}^{(k)} \boldsymbol{m}_{t,\tau}^{(k)T}}\Big) \\ &\quad - \frac{1}{2}\sum_{\tau=t-h}^{t+h} \Big(-2\boldsymbol{\lambda_t}^T \Sigma_t^{-1} \underbrace{\mathbb{E}[\boldsymbol{x_\tau} | \boldsymbol{y_\tau}, \phi_t^{(k)}, \boldsymbol{\lambda_t^{(k)}}]}_{\boldsymbol{m}_{t,\tau}^{(k)}} + \boldsymbol{\lambda_t}^T \Sigma_t^{-1} \boldsymbol{\lambda_t}\Big) \\ &= -\frac{2h+1}{2}\Big(\log\det(\Sigma_t) + \operatorname{trace}\Sigma_t^{-1} R_t^{(k)}\Big) \\ &\quad - \frac{1}{2}\sum_{\tau=t-h}^{t+h} (\boldsymbol{m}_{t,\tau}^{(k)} - \boldsymbol{\lambda_t})^T \Sigma_t^{-1} (\boldsymbol{m}_{t,\tau}^{(k)} - \boldsymbol{\lambda_t}) \end{aligned}$$

The M-step is equivalent to solving a system of $m+1$ non-linear equations, which can be done numerically thanks to the Newton-Raphson algorithm. However, the convergence of the EM algorithm is slow in practice, so Cao et. al. use the EM iterations until the increase of the likelihood function $\ell\Big((\phi_t^{(k)}, \boldsymbol{\lambda_t^{(k)}})|\boldsymbol{y}\Big)$ becomes small, and apply a second order method to achieve convergence [CDVY00]. This method is very heavy though, since a complicated maximization must be carried out on each time window.

8.2.4 The method of routing changes

Consider the problem of estimating the mean $\boldsymbol{x_0}$ of the sequence of traffic matrices $\boldsymbol{x_1}, \ldots, \boldsymbol{x_T}$. We first assume that the routing matrix is A during the whole period of observation. When the link counts $\boldsymbol{y_1}, \ldots, \boldsymbol{y_T}$ are given, a natural approach is to take the least square estimate

$$\operatorname*{argmin}_{\boldsymbol{x}} \quad \sum_{t=1}^{T} \|\boldsymbol{y_t} - A\boldsymbol{x}\|^2 = \operatorname*{argmin}_{\boldsymbol{x}} \quad \left\| \begin{bmatrix} \boldsymbol{y_1} \\ \vdots \\ \boldsymbol{y_T} \end{bmatrix} - \begin{bmatrix} A \\ \vdots \\ A \end{bmatrix} \boldsymbol{x} \right\|^2 .$$

Intuitively, when the number of observations T becomes large, this problem should provide more and more accurate estimations of the mean $\boldsymbol{x_0}$ of the time series of traffic matrices. However, the matrix $[A^T, \ldots, A^T]^T$ involved in the latter problem is rank deficient, because

$$\operatorname{rank}[A^T, \ldots, A^T]^T = \operatorname{rank} A = \operatorname{rank} A^T A \leq l << m,$$

and the problem has an infinity of solutions, which coincide with the solutions of

$$A^T A\boldsymbol{x} = A^T \left(\frac{\sum_{t=1}^{T} \boldsymbol{y_t}}{T} \right).$$

Hence, the problem of estimating the mean $\boldsymbol{x_0}$ is *as ill-posed* as the problem of estimating the whole traffic matrix $X = [\boldsymbol{x_1}, \ldots, \boldsymbol{x_t}]$.

If however the routing matrix is different during each observation period, it is likely that the matrix

$$\mathcal{A} = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_T \end{bmatrix}$$

becomes of full column rank (i.e. $\operatorname{rank} \mathcal{A} = m$). In fact, Soule et. al. [SNC+07] have demonstrated that if the topology of the network is *bidirectional biconnected*, then there always exists an integer T and routing matrices $A_1, \ldots, A_T$ such that $\mathcal{A}$ has full column rank and each routing matrix A_t corresponds to the shortest paths for a set of weights on the links of the network. Soule and his coauthors therefore assumed that the network provider could change the link weights *on purpose*, so that the aggregated routing matrix $\mathcal{A}$ on the global observation period becomes of full rank, and the least square estimation of $\boldsymbol{x_0}$ becomes possible. They further propose a scheme for estimating the variance S of $\overline{\boldsymbol{y}} = [\boldsymbol{y_1}^T, \ldots, \boldsymbol{y_t}^T]^T$ from the sample covariance of the link counts, and suggest to use the Gauss Markov estimator $\hat{\boldsymbol{x}}$ of $\boldsymbol{x_0}$ (cf. Section 2.2.3):

$$\hat{\boldsymbol{x_0}} = (\mathcal{A}^T S^{-1} \mathcal{A})^{-1} \mathcal{A}^T S^{-1} \overline{\boldsymbol{y}}.$$

In fact, the number of routing changes required to let $\mathcal{A}$ be of full rank can be very high. Instead, based on the observation that a small number of flows supports most of the traffic (*elephant and mice* behaviour, 30% of the flows carry 95% of the traffic), and that *elephant* flows have the largest variance, Soule et. al. [SNC+07] have proposed to simply ignore the flows corresponding to the small diagonal terms in the estimated covariance matrix S (by setting them to 0). The number of flows to be estimated is now approximately of $m/3$, and the aggregated routing matrix $\mathcal{A}$ is restricted to the corresponding columns, which can dramatically lower the number of required routing changes.

The same method can be used to estimate a smooth approximation of the traffic: Based on the fact that the traffic is cyclo-stationary with a period of 24 hours, a natural model

for the traffic is:

$$\boldsymbol{x_t} = \boldsymbol{x_0}(t) + \boldsymbol{w_t}$$

where $\boldsymbol{x_0}(t)$ is a deterministic, smooth periodic function (of period 24 hours), and $\boldsymbol{w_t}$ is a centered, stationary random noise process. The authors of [SNC+07] show that the same approach as before can be used to estimate the first Fourier coefficients of $\boldsymbol{x_0}(t)$. To this end, let us approximate $\boldsymbol{x_0}(t)$ by the Fourier expansion

$$\boldsymbol{x_0}(t) = \phi_0(t)\boldsymbol{\theta_0} + \ldots + \phi_{2k}(t)\boldsymbol{\theta_{2k}},$$

where the ϕ_i are the basis cos and sine functions

$$\begin{aligned} \phi_0(t) &= 1 \\ \forall i \in [k], \quad \phi_i(t) &= \cos\left(2\pi i \frac{t}{24}\right) \\ \phi_{k+i}(t) &= \sin\left(2\pi i \frac{t}{24}\right), \end{aligned}$$

where the time t is indicated *in hours*.

The problem is now to estimate the vector of $(2k+1)m$ coefficients $\overline{\boldsymbol{\theta}} = [{\boldsymbol{\theta_0}}^T, \ldots, {\boldsymbol{\theta_{2k}}}^T]^T$ from the observations

$$\boldsymbol{y_t} = A_t \boldsymbol{x_t} = \underbrace{[\phi_0(t)A_t, \ldots, \phi_{2k}(t)A_t]}_{A'_t} \overline{\boldsymbol{\theta}} + \boldsymbol{v_t},$$

where $\boldsymbol{v_t} = A_t \boldsymbol{w_t}$ is a zero-mean stationary random process, whose covariance matrix is $A_t \Sigma A_t^T$, where Σ can be estimated from the link counts [SNC+07]. So we can use the Gauss-Markov estimator

$$\hat{\overline{\boldsymbol{\theta}}} = (\mathcal{A}'^T \Sigma'^{-1} \mathcal{A}')^{-1} \mathcal{A}'^T \Sigma'^{-1} \overline{\boldsymbol{y}},$$

where

$$\mathcal{A}' = \begin{bmatrix} A'_1 \\ \vdots \\ A'_T \end{bmatrix} \quad \text{and} \quad \Sigma' = \begin{pmatrix} A_1 \Sigma A_1^T & & \\ & \ddots & \\ & & A_T \Sigma A_T^T \end{pmatrix}.$$

8.2.5 Spline-based maximum-likelihood estimation

In the previous approach, the number of unknowns (mT) was reduced by considering a temporal basis for the OD flows, which let the vector of parameters of the model ($\overline{\boldsymbol{\theta}}$) be identifiable. Instead, Casas, Vaton, Fillatre and Chonavel have propose a model [CVFC09] in which a spatial basis is assumed: they empirically noticed that when the number of OD flows is large, the sorted components of the vector $\boldsymbol{x_t}$ form a smooth, nondecreasing curve, and that the order of a flow (with respect to the sorted vector of flow volumes) remains stable during long period of times.

Casas et. al. [CVFC09] thus proposed to use a basis $\boldsymbol{s'_1}, \ldots, \boldsymbol{s'_q}$ of cubic spline functions (discretized as vectors with m coordinates) to approximate the smooth curve of the sorted flows, the number of splines q being several order of magnitudes smaller than m. This basis is then rearranged with respect to the order of the flow volumes within a tomogravity estimate $\boldsymbol{x^G}$ of the traffic matrix: the new basis $S = [\boldsymbol{s_1}, \ldots, \boldsymbol{s_q}]$ is such that if i is the index of the k^{th} largest component of $\boldsymbol{x^G}$, then the i^{th} coordinates of $\boldsymbol{s_1}, \ldots, \boldsymbol{s_q}$ are set to the k^{th} coordinate of $\boldsymbol{s'_1}, \ldots, \boldsymbol{s'_q}$, respectively.

Now, since the order of the flows is stable over time, a natural model is

$$\boldsymbol{x_t} = S\boldsymbol{\mu_t} + \boldsymbol{w_t},$$

where $\boldsymbol{w_t}$ is a white Gaussian noise of covariance Σ, and $\boldsymbol{\mu_t}$ is a vector of q coefficients which indicates the importance of each spline basis function at time t. Casas and his coauthors suggest to estimate Σ by using the SNMP data over a short training period and evaluating the sample variances of the tomogravity estimates. The measurement equations can thus be modelled as:

$$\boldsymbol{y_t} = A\boldsymbol{x_t} = AS\boldsymbol{\mu_t} + \boldsymbol{v_t},$$

where $\boldsymbol{v_t} \sim \mathcal{N}(\boldsymbol{0}, A\Sigma A^T)$. Since q is small (typically between 5 and 10), the matrix AS is very likely to have the full column rank property, and the Gauss-Markov estimator of $\boldsymbol{\mu_t}$ is:

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{t}} = \left(S^T A^T (A\Sigma A^T)^{-1} AS\right)^{-1} AS(A\Sigma A^T)^{-1}\boldsymbol{y_t},$$

from which we deduce the spline-based estimator $\hat{\boldsymbol{x}}_{\boldsymbol{t}} = S\hat{\boldsymbol{\mu}}_{\boldsymbol{t}}$. Casas et. al. call this estimator the Spline-based Maximum Likelihood (SML) estimator of $\boldsymbol{x_t}$, because under the Gaussian assumption, the Gauss-Markov estimator above coincides with the maximum-likelihood estimator.

8.3 Estimation based on a few direct measurements

The approaches presented in the previous section (which rely only on the link counts) typically yield an average error of estimation in the order of 20%. Moreover, the error is often huge on certain OD pairs. To overcome this problem, Feldmann et. al. [FGL+01] have proposed a method relying on the network-monitoring tool Netflow which allows to perform direct measurements on the OD flows.

8.3.1 Netflow

Netflow is a network-monitoring tool developed by Cisco [CISb], which collects information for each packet it analyzes. In practice, Netflow aggregates the data to the level of a *flow*, where a flow is defined as a sequence of packets sharing the same source and destination IP address, source and destination port number, IP protocol, interface index

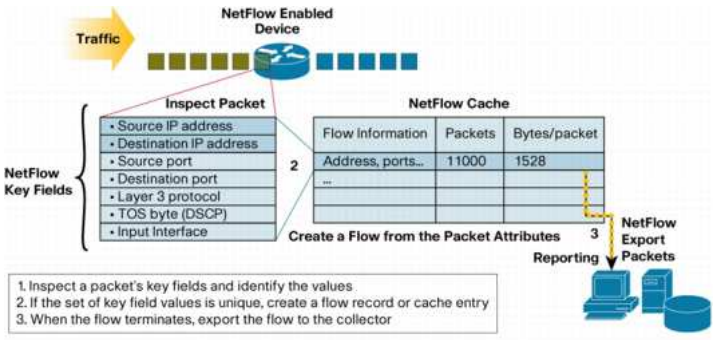


Figure 8.1: Netflow records and exports (Image extracted from Netflow white papers [CISb])

and type of service. The information written on the Netflow records contains, among other things, the source and destination IP address, port numbers, and Autonomous System (AS) numbers, the number of bytes in the flow, a time stamp, and routing information as the IP address of the immediate next-hop. When some information about 20 to 50 flows has been collected, the Netflow record is sent to a global collector (cf. Figure 8.1).

In our problem, we want to use the Netflow records to find the ingress and egress points of the packets in a backbone network. This is not an obvious task, because the IP source and destination are typically connected to several potential ingress and egress routers of the network of interest (cf. Figure 8.2). The solution proposed by Feldmann et. al. [FGL+01] is to activate Netflow directly on all ingress links of the network. In this way, the we can directly infer the ingress node of the packet (it is the place where the packet is being analyzed), and if we dispose of the routing tables, the egress node can be computed by simulating the trajectory of the packet to reach its final IP destination.

However, the limitations of this method were pointed out by Feldmann and his coauthors themselves: in order to measure a complete traffic matrix, one would require to activate Netflow on every ingress link of a network, which may not be practical for several reasons. The quantity of data collected at each router can be huge, which generates storage issues, creates a computational overhead for the router's CPU, and produces heavy records that must be sent through the network to a global collector. Moreover, it is likely that Netflow measurements are not available on each ingress link.

The idea of using Netflow for the estimation of traffic matrices has quickly become obvious after the publication of the latter article. Many authors have proposed some techniques to avoid creating too much overhead with Netflow measurements. This is e.g. the case of Papagiannaki, Taft and Lakhina [PTL04], who formulated recommendations to the developers of Netflow, so that its use could be completely distributed on the Network (so as to reduce the communication overhead). It has also been proposed to use a sampled version of Netflow [FGL+01], which significantly reduce the quantity of data to be analyzed. Choi and Bhattacharyya [CB05] noticed indeed that the overhead involved by Netflow measurements was roughly proportional to the sampling rates.

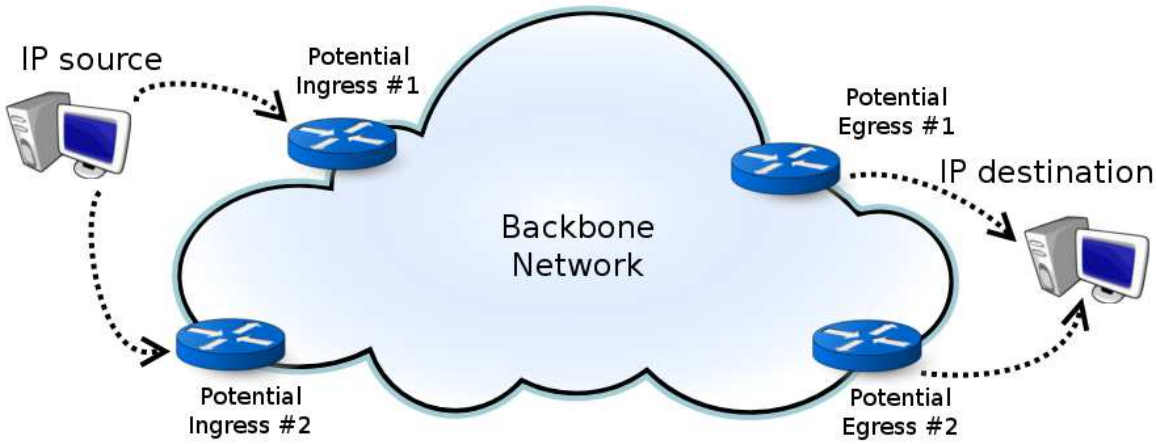


Figure 8.2: There are several possible ingress and egress nodes

For the reasons mentioned above, the intensive use of Netflow (so as to measure each of the m OD flows, and at every time step) is not suitable. Therefore, many authors have proposed to activate Netflow only during a 24 hours-period (so as to calibrate a model), or to measure only a small subset from the m ODs. We next review these methods. We will present a new method to optimize both the deployment of Netflow and the sampling rates in Chapter 10.

8.3.2 Method of fanouts

The method of fanouts was proposed by Papagiannaki, Taft and Lakhina [PTL04]. The key observation at the origin of this method is that, if we consider all the traffic which enters the network at o, the fraction from this traffic that leaves the network at d is very stable over time and exhibits a strong diurnal pattern. More precisely, we define the fanouts

$$f_{o,d}^{(t)} = \frac{x_{o,d}^{(t)}}{\sum_{d'=1}^{n} x_{o,d'}^{(t)}} = \frac{x_{o,d}^{(t)}}{x^{(t),In}(o)}.$$

The authors of [PTL04] have studied the evolution of $f_{o,d}^{(t)}$ over time, and noticed that for almost every OD pair (o,d), $t \mapsto f_{o,d}^{(t)}$ was very stable and periodic (of period 24 hours), e.g. the fanout $f_{o,d}^{(\text{Day1, 3pm})}$ can be used as an accurate estimate for $f_{o,d}^{(\text{Day2, 3pm})}, f_{o,d}^{(\text{Day3, 3pm})}, \ldots$ Note moreover that since the total volume $x^{(t),In}(o)$ of the incoming traffic in o at time t can be measured by the SNMP data, an estimation of the fanouts directly yields an estimation

of the traffic matrix.

The method proposed by the Papagiannaki et. al [PTL04] is thus the following: 24 hours of Netflow measurements are used to compute a baseline of fanouts

$$\left(f_{o,d}^{(t)}\right)_{o\in[n],d\in[n],t\in[24]}$$

(we use time intervals of one hour). The estimation of the traffic matrix at subsequent times is simply carried out by the formula:

$$\hat{x}_{o,d}^{(t)} = f_{o,d}^{(t \bmod 24)} \; x^{(t),In}(o).$$

After a few days, the model needs to be recalibrated, because the fanout baseline is out-of-date. The authors of [PTL04] have proposed a scheme to identify when a recalibration is needed. The results presented in the latter article suggest that 24 hours of Netflow measurements every 4^{th} day allow one to estimate 80% of the traffic matrix with a relative error of estimation below 25%.

A great benefit of this method is that it can be distributed over the network, meaning the estimation of the OD flows can be carried out by the router from which it originates. The number of reports sent through the network to a global collector is thus much smaller and the communication overhead is reduced. Moreover, the recalibration step can be performed independently by each router, thus spreading the measurement effort over space and time.

8.3.3 Principal component analysis

A principal components analysis (PCA) of the (dynamic) traffic matrix X reveals that it may be written as the sum of only a few characteristic *eigenflows*. To see this, Lakhina et. al. [LPC+04] have analyzed the singular value decomposition (SVD) of sample traffic matrices (computed with 3 weeks of Netflow measurements on the European Sprint backbone, or one week of measurement on Abilene). They noticed that the eigenflows could be classified in three categories, namely deterministic flows (accounting for the pseudo-periodic behaviour of the OD flows), spike flows, and noise flows. There is a small number of deterministic flows, which correspond to the largest singular values in the spectrum; the next singular values correspond to spike flows, which describe a sudden and temporary change in the traffic matrix; finally, the lower part of the spectrum mostly contains noise flows (and a few spike flows). Interestingly, Lakhina and his coauthors found that deterministic flows and spike flows capture most energy from the ODs of large volume, while the small flows are dominated by noise. These remarks show that if we can recover the eigenflows corresponding to the upper part of the spectrum, then we shall be able to estimate the largest flows accurately. This led to the PCA approach of Soule et. al. [SLT+05] for the estimation of traffic matrices.

Recall that the rows of X represent the OD flows, while the columns of X are *snapshots* of the traffic matrix during a particular time interval. We consider a a sample traffic matrix X_0 measured with Netflow over a period of T_0 time steps (typically 24 hours), and we let the SVD of X_0^T be

$$X_0^T = USV^T.$$

In the latter expression, U is a matrix with the same dimension as X_0^T ($T_0 \times m$), the columns of which are the *eigenflows* of X; S is a $m \times m$ diagonal matrix and contains the singular values of X, i.e. the i^{th} diagonal element of S indicates the importance of the i^{th} eigenflow in this decomposition. The $m \times m$ square matrix V contains in its i^{th} row the weights that the i^{th} OD assigns to the different eigenflows. Following the structural analysis performed in [LPC+04], the value of X_0 should not change much if we truncate the factor matrices U, S and V to restrict the summations to r eigenflows (with a reasonable value of r), meaning that X_0 may be well approximated by a matrix of rank r:

$$X_0^T \approx U'S'V'^T,$$

where U' is the matrix U restrained to the r principal eigenflows, the matrix S' is the $r \times r$ upper diagonal matrix of S, and V' is formed with the r first columns of V.

By considering a particular snapshot $\boldsymbol{x_t}$ of the traffic matrix (i.e. a row of X_0^T), we obtain:

$$\forall t \in [T_0], \quad \boldsymbol{x_t} \approx V'S'\boldsymbol{u'_t}, \quad \text{and} \quad \boldsymbol{y_t} = A\boldsymbol{x_t} \approx AV'S'\boldsymbol{u'_t},$$

where $\boldsymbol{u'_t}$ denotes the column of U'^T which corresponds to time t. The study of Lakhina et. al. [LPC+04] suggests that the coefficients of S and V are relatively stable over short periods, such that at subsequent times $t > T_0$, the traffic matrix may still be decomposable in this basis, and the coefficients $\boldsymbol{u'_t}$ of the principal largest eigenflows may be well approximated by solving the equation $\boldsymbol{y_t} = AV'S'\boldsymbol{u'_t}$. For small values of r ($r \leq l$), this system of equation is over-determined, and we obtain the least square solution by a pseudo-inverse. Finally, we obtain the following estimate for the traffic matrix at time $t > T_0$:

$$\hat{\boldsymbol{x}}_{\boldsymbol{t}} = V'S'(AV'S')^{\dagger}\boldsymbol{y_t}.$$

In order to avoid negative entries of $\hat{\boldsymbol{x}}_{\boldsymbol{t}}$, the authors of [SLT+05] next set to 0 the negative entries of $\hat{\boldsymbol{x}}_{\boldsymbol{t}}$, and finally use the IPF algorithm (cf. Chapter 9) so that the estimate matches the measurement equations.

Of course, the model needs to be recalibrated after some time, which requires a 24h-period of Netflow measurements. In [SLT+05], the authors propose a test based on the link counts to decide whether a new period of calibration is needed.

8.3.4 Kalman Filter

The idea of using a Kalman Filter for the estimation of traffic matrices was first proposed by Soule et. al. [SLT+05], and further detailed in [SSNT05]. This method assumes the

following linear state-space model:

$$\forall t \in [T], \quad \begin{cases} \boldsymbol{x_{t+1}} & = C\boldsymbol{x_t} + \boldsymbol{w_{t+1}} \\ \boldsymbol{y_t} & = A_t\boldsymbol{x_t} + \boldsymbol{v_t} \end{cases}, \tag{8.3}$$

where the matrix C describes the dynamic of the system; the state-noise process $\boldsymbol{w_t}$ and the measurement-noise process $\boldsymbol{v_t}$ are assumed to be iid Gaussian, centered, and of covariance matrices Q and R, respectively. The problem of finding the minimum variance estimator of the current state $\hat{\boldsymbol{x}}_t$, given a the set of measurements $\{\boldsymbol{y_1}, \ldots, \boldsymbol{y_t}\}$ is classic and can be resolved by a Kalman Filter. We next recall the equations involved by this filter. In what follows, $\hat{\boldsymbol{x}}_{t|t_0}$ represents the estimation of $\boldsymbol{x_t}$ when we use the measurements $\boldsymbol{y_1}, \ldots, \boldsymbol{y_{t_0}}$. The Kalman filter consists in a *prediction step* and a *correction step*, which are applied iteratively. We use the notation $P_{t|t-1} = \mathrm{Var}(\boldsymbol{x_t} - \hat{\boldsymbol{x}}_{\boldsymbol{t|t-1}})$ (resp. $P_{t|t} = \mathrm{Var}(\boldsymbol{x_t} - \hat{\boldsymbol{x}}_{\boldsymbol{t|t}})$) for the *apriori* (resp. *aposteriori*) covariance matrix of $\boldsymbol{x_t}$ at time t.

Prediction Step: Estimation of the new state estimate $\boldsymbol{x_{t+1}}$ based on the information available up to time t.

$$\begin{aligned} \hat{\boldsymbol{x}}_{\boldsymbol{t+1|t}} &= C\hat{\boldsymbol{x}}_{\boldsymbol{t|t}} \\ P_{t+1|t} &= CP_{t|t}C^T + Q \end{aligned}$$

Correction Step: Correction of the estimate based on the new measurement $\boldsymbol{y}_{t+1}$.

$$\begin{aligned} K_{t+1} &= P_{t+1|t}A_{t+1}^T(A_{t+1}P_{t+1|t}A_{t+1}^T + R)^{-1} \\ \hat{\boldsymbol{x}}_{\boldsymbol{t+1|t+1}} &= \hat{\boldsymbol{x}}_{\boldsymbol{t+1|t}} + K_{t+1}(\boldsymbol{y_{t+1}} - A_{t+1}\hat{\boldsymbol{x}}_{\boldsymbol{t+1|t}}) \\ P_{t+1|t+1} &= (\boldsymbol{I} - K_{t+1}A_{t+1})P_{t+1|t} \end{aligned}$$

In order to apply the above Kalman filter for the estimation of the traffic matrix, we need to know the value of the transition matrix C and of the covariance matrices R and Q; we have assumed that the routing matrices $(A_t)_{t\in[T]}$ are available. The authors of [SSNT05] proposed to use 24 hours of Netflow measurements to calibrate the model. The maximum likelihood estimation of C, R and Q can be done with the EM algorithm (see Section 8.2.3). A method to detect when a recalibration of the model is included in [SSNT05].

In a recent paper [CVFC09], Casas, Vaton, Fillatre and Chonavel have proposed an improvement of the latter method. They pointed out that taking the expectation in the first equation of the underlying model (8.3), we obtain $(\boldsymbol{I} - C)\boldsymbol{m_x} = \boldsymbol{0}$, where $\boldsymbol{m_x} = \mathbb{E}[\boldsymbol{x_t}]$ is the average traffic matrix. Therefore, the matrix C must be calibrated in such a way that $\boldsymbol{m_x}$ is in the nullspace of $\boldsymbol{I} - C$, which is certainly not a good model.

To correct this problem, Casas et. al. [CVFC09] assumed that the process has a dynamic mean $\boldsymbol{m_x}(t)$ following the dynamic

$$\boldsymbol{m_x}(t+1) = \boldsymbol{m_x}(t) + \boldsymbol{\zeta_t},$$

where $\boldsymbol{\zeta_t}$ is a zero-mean Gaussian process of covariance Q_ζ. They replaced the transition equation $\boldsymbol{x_{t+1}} = C\boldsymbol{x_t}$ by a transition equation for the variation of the traffic matrix around its mean. To do this, they have defined the new centered augmented state $\boldsymbol{u_t} = \begin{bmatrix} \boldsymbol{x_t} - \boldsymbol{m_x}(t) \\ \boldsymbol{m_x}(t) \end{bmatrix}$, which must follow the dynamic:

$$\forall t \in [T], \quad \left\{ \begin{array}{ll} \boldsymbol{u_{t+1}} & = \begin{pmatrix} C & \\ & \boldsymbol{I} \end{pmatrix} \boldsymbol{u_t} + \begin{bmatrix} \boldsymbol{w_{t+1}} \\ \boldsymbol{\zeta_{t+1}} \end{bmatrix} \\ \boldsymbol{y_t} & = [A_t \ A_t]\boldsymbol{u_t} + \boldsymbol{v_t} \end{array} \right. , \tag{8.4}$$

The matrices C, Q and Q_ζ are assumed to be diagonal and are estimated from a training set of Netflow direct measurements. The measurement equations are assumed exact ($R = 0$). The experiments done by Casas et. al. [CVFC09] show that the Kalman filter based on this centered model with a dynamic mean outperforms that of [SSNT05], and needs less recalibration steps.

8.3.5 Method of Partial Measurements

Contrarily to the previous methods, the approach proposed by Liang, Taft and Yu [LTY06] –which they called PAMTRAM for PArtial Measurements of TRAffic Matrices– does not need a 24 hours-period of intensive measurements. Instead, the proposed scheme is to measure a different subset of OD-pairs during each time interval. Typically, Netflow is activated on only one router during each time interval. Their method can be summarized as follows: during the t^{th} time interval,

- Read both the SNMP data and the direct Netflow measurements from the router that was selected at $t-1$:

$$\boldsymbol{y}_t = A'_t \boldsymbol{x_t}, \quad \text{where } A'_t = \begin{bmatrix} A_t \\ \boldsymbol{e_{N(t),1}}^T \\ \vdots \\ \boldsymbol{e_{N(t),n}}^T \end{bmatrix},$$

  $N(t)$ is the index of the router where Netflow was activated during the t^{th} time interval (as selected at time $t-1$), and $\boldsymbol{e_{o,d}}$ is the canonical vector of the basis of $\mathbb{R}^m$ which has a 1 on the coordinate indexed by the OD pair (o, d). (We recall that we use double indices to facilitate the notation, although the traffic matrices $\boldsymbol{x_t}$ are in vector form.)
- Compute the estimate of the traffic matrix $\hat{\boldsymbol{x}}_{\boldsymbol{t}}$ with the IPF algorithm, which performs the entropic projection of $\hat{\boldsymbol{x}}_{\boldsymbol{t-1}}$ onto the space $\{\boldsymbol{x} : A'_t \boldsymbol{x} = \boldsymbol{y_t}\}$, see Section 9.5.3.

- Choose the Router $N(t+1)$ where Netflow will be activated next. (Several schemes for choosing $N(t+1)$ are presented in [LTY06], some relying on game theoretic arguments.)

Liang, Taft and Yu justify the use of the IPF algorithm (and thus of entropic projections) by the following statistical argument. If we consider the statistical model of Cao et. al. [CDVY00] for $c = 1$ (see Section 8.2.3):

$$\boldsymbol{x_t} \sim \mathcal{N}(\boldsymbol{\lambda}, \Sigma),$$

where $\Sigma = \phi \operatorname{Diag}(\boldsymbol{\lambda})$, then the maximum likelihood estimate of $\boldsymbol{x}_t$ given $\boldsymbol{y}_t = A\boldsymbol{x_t}$ is

$$\mathbb{E}\left[\boldsymbol{x_t}|\boldsymbol{y_t}, \boldsymbol{\lambda}, \phi\right] = \boldsymbol{\lambda} + \Sigma A^T (A\Sigma A^T)^{-1}(\boldsymbol{y_t} - A\boldsymbol{\lambda}). \tag{8.5}$$

The latter expression does not depend on ϕ and corresponds to the solution of the optimization problem

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \sum_{k=1}^{m} \left(\frac{x_k - \lambda_k}{\sqrt{\lambda_k}} \right)^2 \\ \text{s. t.} \quad & A\boldsymbol{x} = \boldsymbol{y_t}. \end{aligned} \tag{8.6}$$

Now, let us consider the optimization problem solved by the IPF algorithm (we refer the reader to Section 9.5.3 for a detailed analysis of this algorithm):

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \sum_{k=1}^{m} x_k \log\left(\frac{x_k}{\lambda_k}\right) \\ \text{s. t.} \quad & A\boldsymbol{x} = \boldsymbol{y_t}. \end{aligned}$$

If the optimum is not too far from $\boldsymbol{\lambda}$, we may use a first order approximation of the cross-entropy criterion:

$$\sum_{k=1}^{m} x_k \log\left(\frac{x_k}{\lambda_k}\right) \approx \sum_{k=1}^{m} x_k \left(\frac{x_k}{\lambda_k} - 1\right).$$

Assuming further that the vector $\boldsymbol{\lambda}$ is nonnegative, and $\sum_{k=1}^{m}(\lambda_k - x_k) \approx 0$, we obtain:

$$\sum_{k=1}^{m} x_k \log\left(\frac{x_k}{\lambda_k}\right) \approx \sum_{k=1}^{m} \frac{x_k^2}{\lambda_k} - x_k + \sum_{k=1}^{m}(\lambda_k - x_k) = \sum_{k=1}^{m} \left(\frac{x_k - \lambda_k}{\sqrt{\lambda_k}} \right)^2,$$

which is the objective function of Problem (8.6).

In practice, $\boldsymbol{\lambda}$ is unknown, but the authors of [LTY06] argue that it can be approximated by the previous estimate of the traffic $\hat{\boldsymbol{x}}_{\boldsymbol{t-1}}$. Therefore, they claim that the entropic projection of $\hat{\boldsymbol{x}}_{\boldsymbol{t-1}}$ onto the space $\{\boldsymbol{x} : A_t'\boldsymbol{x} = \boldsymbol{y_t}\}$ should be a good approximation of the estimator (8.5). The benefit of using the IPF algorithm instead of Formula (8.5) is twofold: the IPF algorithm runs very fast and avoids heavy matrices inversions (as required in (8.5));

the solution returned by the IPF is guaranteed to be positive, which is a desirable feature for any decent estimate of the traffic matrix.

8.4 Brief comparison of the approaches presented in this chapter

We conclude this chapter with a summary of the different approaches to estimate the traffic matrices. Several comparison were done [MTS+02, SLT+05, LTY06, CVFC09] and we try to present them in a unified way (Table 8.1). This table summarizes several properties of the estimates from each method:

- *Average temporal L2 error*: The temporal L2 error of an estimate $\hat{\boldsymbol{x}}_{\boldsymbol{t}}$ of the flows at time t is defined as
$$\frac{\|\boldsymbol{x}_{\boldsymbol{t}} - \hat{\boldsymbol{x}}_{\boldsymbol{t}}\|_2}{\|\boldsymbol{x}_{\boldsymbol{t}}\|_2}.$$
The first column of the table gives the average of this error over the global observation period divided in T time intervals.
- *Average spatial L2 error*: The spatial L2 error of an estimate $\hat{\boldsymbol{x}}_{\boldsymbol{o,d}} = [\hat{x}_{o,d}^{(1)}, \ldots, \hat{x}_{o,d}^{(T)}]^T$ of the time series of the flow volumes from O to D is defined as
$$\frac{\|\boldsymbol{x}_{\boldsymbol{o,d}} - \hat{\boldsymbol{x}}_{\boldsymbol{o,d}}\|_2}{\|\boldsymbol{x}_{\boldsymbol{o,d}}\|_2}.$$
The second column of the table gives the average of this error over the m considered OD flows.
- *OD flows with an error $<$ 20%*: fraction from the m OD flows which have a spatial L2 error lower than 20%.
- *fraction of traffic with an error <10%*: Same as previous column, but the ratio is computed with respect to volume of traffic correctly estimated (instead of the number of flows). This indicator thus gives more weight to heavy flows (by comparison to that of previous column).
- *Netflow measurements*: Do we need Netflow data, and at which frequency ?
- *Adaptivity*: Does the estimate quickly adapt when a sudden change in the traffic matrix occurs ? This feature was analyzed by Soule et. al. [SLT+05]. We think that the adaptivity of the methods of EM, routing changes, and splines must be very bad, because these methods rely on strong assumptions which become wrong when there is a change in the traffic matrix. By contrast, the adaptivity of the PAMTRAM method should be very good, because the inference relies on the tomogravity method, which has an excellent adaptivity.
- *Bias*: Does the mean of the estimate coincide with the mean of the real traffic ?

The entries in the table which are preceded from the sign $\approx$ may not be very accurate, because they were inferred from a (small) graph in [SLT+05].

The data from this table comes from heterogeneous sources (see the superscripts and the notes below the table), and two entries should not be compared on an absolute basis

when they have different superscripts. For example, the upper left entry of the table (25% or 11%) shows that the experimental conditions of [CVFC09] (*natural* symbol $^\natural$), might be easier than in [SLT$^+$05] (dagger symbol †). This indicates that one should not conclude with certainty that the spline method is better than the fanouts method, by comparing the 8% entry and the 15% entry in the first column of the table.

However, the numbers in the table show the order of magnitude of the errors for each method. For example, we see clearly that the methods relying on Netflow produce better results. We see that the tomogravity estimate, which is very simple to compute and does not require any direct measurement, has the best *adaptability*. This estimate can thus be used to track changes in the flow volumes [SLT$^+$05]. Among the methods relying on link counts only, the method based on splines seems to give the best results; among those relying on Netflow, the PAMTRAM approach is probably the most practical (no intensive period of calibration), and gives the best results.

Method	Avg temporal L2 error	Avg spatial L2 error	OD flows with error <20%	fraction of traffic with an error < 10%	Netflow measurements	Adaptivity	Bias
Tomogravity [ZRDG03]	25%[†] or 11%[♮]	≈27%[†]	25%[†]	40%[♮]	None	Excellent[†]	Strong[†] (+ or −)
EM algorithm [CDVY00]		22%[♭]	52%[♭]		None	Very Bad ?	
Routing Changes [SNC+07]	45%[†]	≈43%[†]	4%[†]		No, but several routing changes	Very Bad ?	
Splines ML [CVFC09]	8%[♮]			75%[♮]	No, calibration with 1h of SNMP data	Very Bad ?	
Fanouts [PTL04]	15%[†]	≈18%[†]	63%[†]		24 hours every $\approx 4^{\text{th}}$ day	Bad[†]	No bias[†]
PCA [SLT+05]	12%[†]	≈16%[†]	47%[†]		24 hours every $\approx 4^{\text{th}}$ day	Bad[†]	Negative bias[†]
Kalman [SSNT05]	10%[†]	≈16%[†] or 21%[‡]	33%[†]	65%[‡]	24 hours every $\approx 4^{\text{th}}$ day	Bad[†]	Negative bias[†]
centered Kalman [CVFC09]	4.5%[♮]			65%[♮]	24 hours every $\approx 10^{\text{th}}$ day	Bad ?	No bias[♮]
PAMTRAM [LTY06]		16.5%[‡]		92%[‡]	Continuous, on only one router	very good ?	

Table 8.1: Summary comparison of the methods for the estimation of Internet traffic matrices

References:

[♭]:Comparison of the Bayesian approaches with a LP approach [MTS+02]. Synthetic data on a 14-nodes topology.

[†]: Comparative study of [SLT+05]. Data from the Sprint Network, aggregated to the level of 13 PoPs. Only the largest flows, representing 95% of the total traffic, are taken into account.

[‡]: Comparison between the PAMTRAM and the Kalman approach [LTY06]. Data from the Sprint Network, aggregated to the level of 12 PoPs. Only the largest flows, representing 90% of the total traffic, are taken into account.

[♮]: Article [CVFC09]. Data Of Abilene (12 nodes) available at [Abi].

Chapter 9

Information theory and entropic projections

In this chapter, we review the information theoretic approach to the problem of inferring the traffic matrix from link counts, which leads to entropy minimization problems with linear constraints. We shall study the latter optimization problem in detail, and in particular the similarities with the classic problem of *matrix balancing*.

9.1 The gravity model

Let us consider a network with a set of n sources and n' sinks. An Internet provider wishes to infer the traffic matrix (with $m = nn'$ unknowns), but the only piece of information at her disposal is the volume of traffic $(t_1^{In}, \ldots, t_n^{In})$ on the n ingress links and the traffic $(t_1^{Out}, \ldots, t_{n'}^{Out})$ on the n' egress links. This problem is actually equivalent to the problem of inferring the traffic matrix from link counts (cf. Chapter 8), for the star-shaped network depicted on Figure 9.1, in which the small square in the middle is a *black box* accounting for all the unknown internal behaviour of the network.

We can normalize the vector of traffic $\boldsymbol{x}$ so that it sums to 1: $p_i := \frac{x_i}{\sum_j x_j}$ represents the probability that a packet travelling on the network belongs to the i^{th} OD pair. Following the principle of maximum entropy, the probability distribution which best represents the current state of knowledge is, among all those distributions satisfying the measurement equations, the one with largest entropy. We illustrate this postulate with a classic combinatorial argument. Assume that the total number of packets that have travelled on the network during the considered period is $N = \sum_{i=1}^{n} t_i^{In} = \sum_{i=1}^{m} x_i$. We can count the number of allocations of these N packets to the $m = nn'$ OD pairs for which the traffic is $\boldsymbol{x} = (x_1, \ldots, x_m)^T$. This number is given by the multinomial coefficient

$$W(\boldsymbol{x}) = \frac{N!}{x_1! \cdots x_m!}.$$

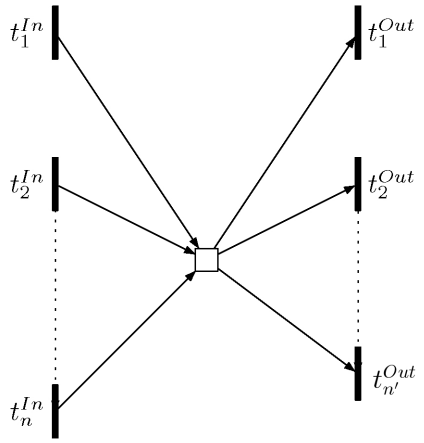


Figure 9.1: The star-shaped Network

In absence of additional information, we are going to assume that every one of these allocations has an equal probability of outcome, and we are going to select the allocation $\boldsymbol{x}$ which is the most likely to be observed. In other words, our estimate of $\boldsymbol{x}$ must maximize $W(\boldsymbol{x})$, subject to the observations $\boldsymbol{t^{In}}$ and $\boldsymbol{t^{Out}}$ on the access links of the network. Now, $W(\boldsymbol{x})$ has a complex expression, which is defined only for integer values of $\boldsymbol{x}$. We remedy this problem by maximizing $\frac{1}{N}\log W(\boldsymbol{x})$ instead, and by taking the limit as N grows to infinity:

$$\frac{1}{N}\log W(\boldsymbol{x}) = \frac{1}{N}(\log N! - \sum_{i=1}^{m}\log(Np_i)!),$$

where we have set $p_i = \frac{x_i}{N}$. We assume that $p_i > 0$ for every i (which is almost sure when we let $N \to \infty$), and the Stirling approximation yields

$$\begin{aligned}\lim_{N\to\infty}\frac{1}{N}\log W(\boldsymbol{x}) &= \frac{1}{N}(N\log N - \sum_i (Np_i)\log Np_i)\\ &= \log N - \sum_i p_i \log N - \sum_i p_i \log p_i\\ &= -\sum_i p_i \log p_i.\end{aligned}$$

In conclusion, we are going to select the distribution of the traffic which maximizes the entropy $H(\boldsymbol{x}) = -\sum_{i=1}^{m}\frac{x_i}{N}\log\frac{x_i}{N}$. We now use double indices, such that x_{od} represents the traffic from o to d in order to simplify the notation. Our estimate will be the solution of

the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{x}\geq\mathbf{0}} \quad & \sum_{o=1}^{n}\sum_{d=1}^{n'} \frac{x_{od}}{N}\log\frac{x_{od}}{N} \\ \text{s.t.} \quad & \sum_{d=1}^{n'} x_{od} = t_o^{In} \quad (\forall o \in [n]) \\ & \sum_{o=1}^{n} x_{od} = t_d^{Out} \quad (\forall d \in [n']) \end{aligned} \tag{9.1}$$

Interestingly, this optimization problem has a closed-form solution which we derive below. Let us form the Lagrangian

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda^{In}}, \boldsymbol{\lambda^{Out}}) = \sum_{o=1}^{n}\sum_{d=1}^{n'} \frac{x_{od}}{N}\log\frac{x_{od}}{N} + \sum_{o=1}^{n}\lambda_o^{In}(t_o^{In} - \sum_{d=1}^{n'} x_{od}) + \sum_{d=1}^{n'}\lambda_d^{Out}(t_d^{Out} - \sum_{o=1}^{n} x_{od}).$$

Let $\boldsymbol{x}$ be a solution of Problem (9.1). Since the objective function of this problem is convex, and the constraints are affine, there must exist a Lagrange multiplier $\boldsymbol{\lambda}$ such that $(vecx, \boldsymbol{\lambda})$ is a saddle point of the Lagrangian:

$$\forall o \in [n],\ \forall d \in [n'],\ \frac{1}{N}(\log\frac{x_{od}}{N} + 1) = \lambda_o^{In} + \lambda_d^{Out}.$$

Setting $\boldsymbol{u} = \exp(N\boldsymbol{\lambda^{In}})$ and $\boldsymbol{v} = N\exp(\boldsymbol{N\lambda^{Out}} - \mathbf{1})$, where the exponential is taken component-wise, we see that x_{od} is of the form $u_o v_d$ for some vectors $\boldsymbol{u}$ and $\boldsymbol{v}$. Substituting this expression to $\boldsymbol{x}$ in the constraints of Problem (9.1), we see that $\boldsymbol{u}$ must be proportional to $\boldsymbol{t^{In}}$ and $\boldsymbol{v}$ to $\boldsymbol{t^{Out}}$. Since the total traffic must sum to N, we finally obtain:

$$\forall o \in [n],\ \forall d \in [n'], \quad x_{od} = \frac{t_o^{In} t_d^{Out}}{N}. \tag{9.2}$$

This estimate of the traffic matrix, in which the traffic from o to d is proportional to the incoming traffic at o, multiplied by the outgoing traffic at d is traditionally referred as the *gravity* estimate, because of the similarity to Newton's gravity law. Also note that this is a rank-one approximation of the traffic matrix.

9.2 Entropic projections

The previous section justifies the use of the gravity model as a prior for the traffic matrix. According to the information theory, a natural approach is to take as an estimate the distribution of flows which satisfies all the measurement equations (access &internal link counts), and is as hard to discriminate from the prior as possible (Principle of Minimum Discrimination Information). We thus want to choose, among the vector of flows satisfying the measurements, the one which is the closest to the prior, in terms of Kullback-Leibler

divergence, (a.k.a. cross entropy). Here again, a combinatorial argument similar to the one given in previous section can justify this approach. If we assume that every one of the N packets is allocated to an OD with respect to the prior probability, then the vector $\boldsymbol{x}$ which minimize the Kullback Leibler divergence (with respect to the prior) corresponds to the most likely allocation of the packets to the OD pairs. This method has often been used in (road) transportation planning, see e.g. Levinson and Kumar [LK94] and in telecommunication networks [CDVY00, ZRDG03, LTY06], where the entropic projection of the gravity prior onto the space of flows which satisfy the measurement equations is called *tomogravity* estimate.

In the remaining of this chapter, we denote the prior estimate of the flows by $\boldsymbol{c} = (c_1, \ldots, c_m)^T$. This prior can be e.g. the gravity estimate, or an estimation of the flows at a previous point in time, or a combination of both. The entropy projection problem follows:

$$\begin{aligned} \min_{\boldsymbol{x} \geq \boldsymbol{0}} \quad & \sum_{i \in [m]: c_i > 0} x_i (\log\left(\frac{x_i}{c_i}\right) - 1) \\ \text{s.t.} \quad & A\boldsymbol{x} = \boldsymbol{y} \\ & \boldsymbol{x} \geq 0 \\ & c_i = 0 \Longrightarrow x_i = 0 \end{aligned} \qquad (9.3)$$

We will assume that all the c_i are positive in further discussion, because we only need solve the problem (9.3) for the indices i such that $c_i \neq 0$. The constraint $A\boldsymbol{x} = \boldsymbol{y}$ represents the measurement equations (7.1), which comprise the SNMP data and Netflow measurements (if any). In fact, we have subtracted $\sum_i x_i$ from the expression of cross-entropy to simplify subsequent calculations (the resulting expression $D(\boldsymbol{x}\|\boldsymbol{c}) = \sum_{i=1}^m x_i(\log\left(\frac{x_i}{c_i}\right) - 1)$ is still referred as the Kullback Leibler divergence). This does not change the value of the minimizer $\boldsymbol{x}$, since the total traffic $\sum_i x_i$ is constrained by the observation equations $A\boldsymbol{x} = \boldsymbol{y}$:

$$\text{Assumption } (A.1) \qquad\qquad A\boldsymbol{x} = \boldsymbol{y} \text{ implies } \quad \sum_{i=1}^{m} x_i = N.$$

In a variety of problems, the $q \times m$ matrix A only has $0/1$ entries (when the traffic of an OD pair is never split among different routes). In this case, we will often use the following notation and terminology, which suggest that only SNMP measurements (link counts) are available, and that A is the $(0/1)$-routing matrix of the network: we shall use the index e (as edge) to denote a particular measurement, and the index p (as pair) to denote a particular OD pair. We use the notation $e \in p := \{e \in [q] : A_{ep} = 1\}$ and $p \ni e := \{p \in [m] : A_{ep} = 1\}$. This notation is better understood in the context of link measurements, where $e \in p$ is the set of all links that belong to the path of the pair p, while $p \ni e$ represents the set of all the OD pairs p that traverse the link e. With this notation, the measurement equation $A\boldsymbol{x} = \boldsymbol{y}$ becomes:

$$\forall e \in [q],\ y_e = \sum_{p \ni e} x_p.$$

9.2.1 The dual problem

Let us denote the Lagrange multiplier associated with the constraints of Problem (9.3) as $\boldsymbol{\lambda}$. We can form the Lagrangian:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \;=\; \sum_{i\in[m]} (x_i \log(\frac{x_i}{c_i}) - x_i) + \boldsymbol{\lambda}^T(\boldsymbol{y} - A\boldsymbol{x}) \tag{9.4}$$

The objective function of Problem (9.3) is strictly convex, and the constraints are affine functions. Therefore, strong duality holds and finding a solution $\boldsymbol{x}$ to Problem (9.3) is equivalent to finding some Lagrange multiplier $\boldsymbol{\lambda}$ such that $(\boldsymbol{x}, \boldsymbol{\lambda})$ is a saddle point.

$$\begin{aligned} \forall i \in [m], \quad & \frac{\partial}{x_i}(\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda})) = 0 \\ \iff \quad & \log(\frac{x_i}{c_i}) = (A^T\boldsymbol{\lambda})_i \end{aligned}$$

We summarize the latter necessary condition in vector notation:

$$\boldsymbol{x} = \boldsymbol{c} \odot \exp(A^T\boldsymbol{\lambda}), \tag{9.5}$$

where the multiplication $\odot$ is elementwise, as well as the exponential. The latter expression makes it possible to give the Lagrange dual function $g(\boldsymbol{\lambda}) = \min_{\boldsymbol{x}\in\mathbb{R}^m} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda})$ in closed form, and the dual of Problem (9.3) is the unconstrained maximization problem:

$$\sup_{\boldsymbol{\lambda}\in\mathbb{R}^q} \; g(\boldsymbol{\lambda}) := \boldsymbol{y}^T\boldsymbol{\lambda} - \boldsymbol{c}^T \exp(A^T\boldsymbol{\lambda}) \tag{9.6}$$

The first order optimality condition of this problem $\nabla g(\boldsymbol{\lambda}) = \mathbf{0}$ is:

$$\boldsymbol{y} = \underbrace{A\Big(\boldsymbol{c} \odot \exp(A^T\boldsymbol{\lambda})\Big)}_{S(\boldsymbol{\lambda})}. \tag{9.7}$$

We now make the following (weak) assumption, which makes possible to express the latter optimality condition as a system of polynomial equations:

Assumption $(A.2)$ — A has only rational entries, i.e.

$$\exists \beta \in \mathbb{N},\ \exists [\alpha_{i,p}] \in \mathbb{Z}^{q\times m} : \quad \forall i, j,\ A_{i,p} = \frac{\alpha_{i,p}}{\beta}$$

Under this assumption, the components of $S(\boldsymbol{\lambda})$ can be rewritten as:

$$\forall i \in [q], \quad S_i(\boldsymbol{\lambda}) = \sum_{p=1}^{m} A_{i,p} c_p \exp(\beta^{-1} \sum_{j=1}^{q} \alpha_{j,p}\lambda_j).$$

We define $\boldsymbol{u} = \exp(\beta^{-1}\boldsymbol{\lambda})$, so that the latter expression is polynomial in $\boldsymbol{u}$:

$$\forall e \in [q], \quad S_i(\boldsymbol{\lambda}) = P_i(\boldsymbol{u}) = \sum_{p=1}^{m} A_{i,p} c_p \prod_{j=1}^{q} u_j^{\alpha_{j,p}}.$$

The polynomial application P maps $(\mathbb{R}_+)^q$ onto itself, and the optimality condition of Problem (9.3) are:

$$P(\boldsymbol{u}) = P(\exp(\beta^{-1}\boldsymbol{\lambda})) = S(\boldsymbol{\lambda}) = \boldsymbol{y}. \tag{9.8}$$

We point out that if A has only $0/1$ entries, then the polynomial P takes a simple form which is linear in each variable (cf. Example 9.2.1).

Proposition 9.2.2. *If the equation $P(\boldsymbol{u}) = \boldsymbol{y}$ has a positive solution $\boldsymbol{u}$, or equivalently if Equation* (9.7) *has a solution $\boldsymbol{\lambda}$, we obtain the unique solution $\boldsymbol{x}$ of Problem* (9.3) *by setting $\boldsymbol{x} = \boldsymbol{c} \odot \exp(A^T\boldsymbol{\lambda})$.*

Proof. The Lagrangian of this problem is non-differentiable for vectors $\boldsymbol{x}$ lying on the boundary of the positive cone $\partial(\mathbb{R}^+)^m$. So let us assume that the solution of Problem (9.3) is positive. In this case, as the affine constraints are automatically qualified, the duality gap vanishes and the problem of finding a optimal Lagrange multiplier $\boldsymbol{\lambda}$ becomes equivalent to solving (9.3). Conversely, if the optimal multiplier is $\boldsymbol{\lambda}$, then the primal solution $\boldsymbol{x} = \boldsymbol{c} \odot \exp(A^T\boldsymbol{\lambda})$ is positive. □

9.3 Existence and uniqueness results

In this section, we will explicit necessary and sufficient conditions so that the dual problem (9.6) has a solution and is unique.

We denote by J the Kullback Leibler divergence appearing in the objective function of Problem (9.3) :

$$J(\boldsymbol{x}) = \sum_{i=1}^{m} x_i(\log(\frac{x_i}{c_i}) - 1). \tag{9.9}$$

Proposition 9.3.1. *The* supremum *in Problem* (9.6) *is attained by a vector $\boldsymbol{\lambda}^* \in \mathbb{R}^q$ if and only if the equation $A\boldsymbol{x} = \boldsymbol{y}$ has a solution $\boldsymbol{x}^{\mathbf{0}} > \mathbf{0}$.*

Example 9.2.1. When the matrix A has only $0/1$, we can simplify the form of the polynomial equations (9.8), by using the previously introduced notation:

$$P_e(\boldsymbol{u}) = \sum_{p \ni e} c_p \prod_{e' \in p} u_{e'} = y_e.$$

Let us consider the toy network of Example 8.2.1. If the prior estimate is $\boldsymbol{c} = (c_{1,2}, c_{2,3}, c_{1,3})^T$ and the vector of link counts is $\boldsymbol{y} = (y_a, y_b)^T$, the system of polynomials associated to Problem (9.3) reads:

$$\begin{aligned} c_{1,2}u_a + c_{1,3}u_a u_b &= y_a \\ c_{2,3}u_b + c_{1,3}u_a u_b &= y_b. \end{aligned}$$

Proof. We assume that there exists a vector $\boldsymbol{x^0} > \mathbf{0}$ such that $\boldsymbol{y} = A\boldsymbol{x^0}$. Let $\mathcal{F}$ denote the feasible set $\{\boldsymbol{x} \geq \mathbf{0} : \ A\boldsymbol{x} = \boldsymbol{y}\}$. This set is nonempty (it contains $\boldsymbol{x^0}$), closed, and bounded by Assumption (A.1). Therefore, the strictly convex function J admits a unique minimizer $\boldsymbol{x}^*$ on $\mathcal{F}$. We shall now see that $\boldsymbol{x}^*$ is not on the boundary $\partial(\mathbb{R}_+)^q$ of the positive cone, i.e. $\boldsymbol{x}^* > \mathbf{0}$.

Let $\mathcal{I}$ be the set of all indices i such that $x_i^* = 0$. If $\mathcal{I}$ is not empty, then $\boldsymbol{x}^* \neq \boldsymbol{x}^0$. Let t be in $[0; 1]$; we define $\boldsymbol{x^t} = (1-t)\boldsymbol{x}^* + t\boldsymbol{x^0}$. Clearly, $\boldsymbol{x^t} \in \mathcal{F}$. Now, let Φ be the function:

$$\Phi(t) = J(\boldsymbol{x^t}) = \sum_{i\in\mathcal{I}} tx_i^0 \log\left(\frac{tx_i^0}{c_i}\right) + \sum_{j\notin\mathcal{I}}((1-t)x_j^* + tx_j^0)\log\left(\frac{(1-t)x_j^* + tx_j^0}{c_j}\right) - \underbrace{\sum_{k=1}^{m} x_k^t}_{N}.$$

We have $\Phi(0) = J(\boldsymbol{x}^*)$, and for all $t > 0$,

$$\Phi'(t) = \sum_{i\in\mathcal{I}} x_i^0(\log\left(\frac{tx_i^0}{c_i}\right) + 1) + \sum_{j\notin\mathcal{I}}(x_j^0 - x_r^*)(\log\left(\frac{(1-t)x_j^* + tx_j^0}{c_j}\right) + 1).$$

One can easily verify that

$$\mathcal{I} \neq \emptyset \Longrightarrow \lim_{t\to 0^+} \Phi'(t) = -\infty.$$

Hence, if $t > 0$ is small enough,

$$\begin{array}{rcl} & \frac{\Phi(t)-\Phi(0)}{t} & < 0 \\ \iff & \Phi(t) - \Phi(0) & < 0 \\ \iff & J(\boldsymbol{x^t}) & < J(\boldsymbol{x}^*). \end{array}$$

This is in contradiction with $\boldsymbol{x}^*$ being the minimum of J over $\mathcal{F}$. So, $\mathcal{I} = \emptyset$ and $\boldsymbol{x}^* > \mathbf{0}$. Hence, the Lagrangian is differentiable at $\boldsymbol{x}^*$ and since the primal problem (9.3) is strictly feasible, we know from the strong duality theorem that the dual problem (9.6) attains its solution. In other words, there exists an optimal Lagrange multiplier $\boldsymbol{\lambda}^*$ such that

$$S(\boldsymbol{\lambda}^*) = \boldsymbol{y}.$$

Conversely, if there is a vector $\boldsymbol{\lambda}^* \in \mathbb{R}^q$ such that $S(\boldsymbol{\lambda}^*) = \boldsymbol{y}$, it is clear that the vector $\boldsymbol{x}^* = \boldsymbol{c} \odot \exp(A^T\boldsymbol{\lambda}^*) > 0$ is a solution of the measurement equation $\boldsymbol{y} = A\boldsymbol{x}$. □

Remark 9.3.1. If there is no noise in the measurements, then the real traffic $\boldsymbol{x}$ is a solution of the equation $A\boldsymbol{x} = \boldsymbol{y}$. If $\boldsymbol{x} > \mathbf{0}$, i.e. there is some traffic on all the OD pairs, then the condition of Proposition 9.3.1 is fulfilled and the dual problem has a solution.

We next present a necessary and sufficient condition which ensures that the solution of Problem (9.6) is unique.

Proposition 9.3.2. *Let g be the dual function defined as in* (9.6).

1. $\forall \boldsymbol{\lambda} \in \mathbb{R}^q, \nabla^2 g(\boldsymbol{\lambda})$ *is a positive semidefinite symmetric matrix (and hence g is convex).*

2. *Furthermore, this Hessian is positive definite if and only if A has full row-rank (i.e. the rows of A are linearly independent). In that case, g is strictly convex and the minimizer (if it exists) is unique.*

Proof. The Hessian $\nabla^2 g(\boldsymbol{\lambda})$ is the matrix whose entry (i,j) equals $\frac{\partial^2 g(\boldsymbol{\lambda})}{\partial\lambda_i \partial\lambda_i} = \frac{\partial S_i(\boldsymbol{\lambda})}{\partial\lambda_j}$:

$$\Big(\nabla^2 g(\boldsymbol{\lambda})\Big)_{i,j} = \sum_{p=1}^{m} A_{i,p}\ A_{j,p}\ c_p\ \exp(A^T\boldsymbol{\lambda})_p.$$

Using a matrix notation, one may easily verify that $\nabla^2 g(\boldsymbol{\lambda}) = ADA^T$, where D is the diagonal matrix $\mathrm{Diag}\left(\boldsymbol{c}\odot\exp(A^T\boldsymbol{\lambda})\right) \succ 0$. Hence, it is elementary that $\nabla^2 g(\boldsymbol{\lambda})$ is positive semidefinite, and positive definite if and only if A has full row-rank. □

In fact, if the rows of A are not linearly independent, then it means that there is some redundancy in the measurement equations. Hence, in the case of noiseless measurements, we do not loose any information by removing those rows in A that are linear combinations of the others. For the remaining of this chapter, we thus make the following assumption:

$$\text{Assumption } (A.3) \qquad \text{The rows of } A \text{ are linearly independent}$$

Corollary 9.3.3. *The application P is a $\mathcal{C}^1-$diffeomorphism which maps $(\mathbb{R}_+^*)^q$ onto the cone $K := \{\boldsymbol{y}\in\mathbb{R}^q :\ \exists \boldsymbol{x} > \mathbf{0} :\ A\boldsymbol{x} = \boldsymbol{y}\}$. In particular, P^{-1} exists on K, and is one to one.*

Proof. Let $\boldsymbol{y}\in K$. By Proposition 9.3.1, the dual problem (9.6) has a solution $\boldsymbol{\lambda}^*\in\mathbb{R}^q$. which is unique by strict convexity of g (Proposition (9.3.2) and Assumption $(A.3)$). Hence, $\boldsymbol{u}^* = \exp(\beta^{-1}\boldsymbol{\lambda}^*)$ is the unique solution of the polynomial system $P(\boldsymbol{u}) = \boldsymbol{y}$. This shows already that P^{-1} exists and is one-to-one.

Moreover, we have from the chain rule:

$$\nabla^2 g(\boldsymbol{\lambda}) = \frac{\partial S(\boldsymbol{\lambda})}{\partial(\lambda_1,\dots,\lambda_q)} = \frac{\partial P(\exp(\boldsymbol{\beta^{-1}\lambda}))}{\partial(\lambda_1,\dots,\lambda_q)}\,\mathrm{Diag}\left(\beta^{-1}\exp(\beta^{-1}\boldsymbol{\lambda})\right),$$

where $\frac{\partial}{\partial(\lambda_1,\dots,\lambda_q)}$ denotes the Jacobian matrix. After the change of variable $\boldsymbol{u} = \exp(\beta^{-1}\boldsymbol{\lambda})$, we obtain:

$$\forall \boldsymbol{u} > \mathbf{0}, \frac{\partial P(\boldsymbol{u})}{\partial(\lambda_1,\dots,\lambda_q)} = \beta\ \nabla^2 g(\lambda)\,\mathrm{Diag}(\boldsymbol{u})^{-1}.$$

By Proposition (9.3.2) and Assumption $(A.3)$, the Hessian matrix $\nabla^2 g(\lambda)$ is positive definite, such that the Jacobian determinant

$$\left|\frac{\partial P(\boldsymbol{u})}{\partial(\lambda_1,\dots,\lambda_q)}\right|$$

is positive for all $\boldsymbol{u} > \mathbf{0}$. Finally, $P : (\mathbb{R}_+^*)^q \longrightarrow K$ is $\mathcal{C}^1$, injective, and its Jacobian determinant never vanishes. Hence, the statement of the corollary follows from the global inverse mapping theorem. □

9.4 Historic relation with Matrix balancing

The problem of *Matrix Balancing* has been widely studied in the 60's and 70's by several authors [Bru68, Men67, MS69, Bre67a, Bre67b]. The problem (9.3) of minimizing an entropy under affine constraints can actually be considered as a generalization of a *Matrix Balancing*. For this reason, most of the algorithms to solve (9.3) are generalizations of algorithms that have been used to solve the *Matrix Balancing* problem. In this section, we give a brief review on this problem and the algorithms to solve it.

9.4.1 The Matrix Balancing problem

Given a matrix H of size $n \times n'$ with nonnegative entries, a row vector $\overline{\boldsymbol{c}}$ of size n' and a column vector $\overline{\boldsymbol{r}}$ of size n, the problem of *matrix balancing* is that of finding the matrix X the closest to H (in terms of Kullback-Leibler distance), whose row sums are given by $\overline{\boldsymbol{r}}$ and whose column sums are given by $\overline{\boldsymbol{c}}$:

$$\begin{aligned} \min_{X \in \mathbb{R}^{n \times n'}} \quad & \sum_{\substack{i \in [n] \\ j \in [n']}} X_{i,j} \log\left(\frac{X_{i,j}}{H_{i,j}}\right) \\ \text{s.t.} \quad & \forall i \in [n], \sum_{j \in [n']} X_{ij} = \overline{r}_i \\ & \forall j \in [n'], \sum_{i \in [n]} X_{ij} = \overline{c}_j \\ & X_{i,j} \geq 0. \end{aligned} \tag{9.10}$$

In fact, this problem is strictly equivalent to Problem (9.3), when the graph is the star-shaped network introduced in Section 9.1 (cf. Figure 9.1), with $\overline{\boldsymbol{r}} = \boldsymbol{t}^{\boldsymbol{In}}$ and $\overline{\boldsymbol{c}} = \boldsymbol{t}^{\boldsymbol{Out}}$.

9.4.2 Algorithms for Matrix balancing

The first algorithm to solve Problem (9.10) was attributed to Sheleikhovskii (1930's) by Bregman [Bre67a], who further proved the convergence of this method. This Algorithm has been called *matrix scaling*, because each iteration of the algorithm consists either in a normalization of the rows or of the columns of X (see Algorithm 9.4.1).

Algorithm 9.4.1 Matrix scaling algorithm of Bregman

$X^{(0)} \leftarrow H$
for $t = 0, 1, 2, \ldots,$ **do**
$\forall (i,j)\ X_{ij}^{(t^+)} \leftarrow X_{ij}^{(t)} \ \frac{\overline{r}_i}{\sum_{j'} X_{ij'}^{(t)}}$
$\forall (i,j)\ X_{ij}^{(t+1)} \leftarrow X_{ij}^{(t^+)} \ \frac{\overline{c}_j}{\sum_{i'} X_{i'j}^{(t^+)}}$
end for

Interestingly, with the help of Hilbert's projective metric and Perron Frobenius theory [FL89] (see also [BR97]), it was shown that this algorithm has a linear rate of convergence (for the Hilbert metric), and that the rate of convergence is bounded by $\tanh\frac{\Delta(H)}{4}$, where $\Delta(H)$ is the diameter of the image of $(\mathbb{R}^+)^n$ under the operator H:

$$\Delta(H) = \log \max_{i,j,k,l} \ \frac{H_{ik} H_{jl}}{H_{il} H_{jk}}.$$

Taking into account the fact that the optimal solution X^* can be obtained only by rows and columns scalings, Brualdi [Bru68] looked for a solution of the form $X^* = U^* H V^*$, where U^* and V^* are diagonal matrices. He proved that such a diagonal scaling was possible when H is fully indecomposable under an simple condition on its zero pattern. He further proposed the update rules of Algorithm 9.4.2 to compute $U^* = \mathrm{Diag}(\boldsymbol{u}^*)$ and $V^* = \mathrm{Diag}(\boldsymbol{v}^*)$.

Algorithm 9.4.2 Dual matrix scaling algorithm

$\boldsymbol{u^{(0)}} \leftarrow \mathbf{1} \in \mathbb{R}^n$
$\boldsymbol{v^{(0)}} \leftarrow \mathbf{1} \in \mathbb{R}^{n'}$
for $t = 0, 1, 2, \ldots,$ **do**
$\forall i \in [n],\ u_i^{(t+1)} \leftarrow \frac{\overline{r}_i}{\sum_j H_{ij}\ v_j^{(t)}}$
$\forall j \in [n'],\ v_j^{(t+1)} \leftarrow \frac{\overline{c}_j}{\sum_i H_{ij}\ u_i^{(t)}}$
$X^{(t+1)} \leftarrow \mathrm{Diag}(\boldsymbol{u^{(t+1)}})\ H\ \mathrm{Diag}(\boldsymbol{v^{(t+1)}})$
end for

Algorithm 9.4.2 is basically a dual approach to solve problem (9.10), and the sequence of matrices $(X^{(t)})_{t\in\mathbb{N}}$ that it generates is exactly the same as the one generated by Algorithm 9.4.1. Menon and Schneider [Men67, MS69] studied the spectrum of the operator T that associates $\boldsymbol{v^{(t)}}$ to $\boldsymbol{v^{(t+1)}}$, and showed that it has a single eigenvalue, namely 1 in the natural case where $\sum_i \overline{r}_i = \sum_j \overline{c}_j$: $\boldsymbol{v^*}$ is therefore a fixed-point of T. This dual approach allows to store much less variables than the primal one.

9.5 Algorithms for the problem of entropic projection

In this section, we shall review the algorithms to solve Problem (9.3), with a particular focus on the relations with algorithms for matrix balancing. We show that the direct

generalization of Algorithm 9.4.2 works if and only if all the OD pairs considered in the network are of length at most 2. We next present a variant of the latter algorithm, in which the coordinates of the variable are updated one at a time, in a cyclic manner (instead of being updated simultaneously). This algorithm is called *Iterative Proportional Fitting* (IPF) in the traffic matrix literature [CDVY00, LTY06], and belongs to the class of cyclic projection algorithms. Therefore it has a linear rate of convergence.

9.5.1 A fixed point algorithm

We have seen in the previous section that the matrix balancing algorithm was well suited to solve Problem (9.3) on the star-shaped network (Figure 9.1). Assume that the prior estimation of the OD traffic is $\boldsymbol{c} = (c_{i,j})_{i\in[n],j\in[n']}$. On this network, the polynomial equation $P(\boldsymbol{u}) = \boldsymbol{y}$ (9.8) takes the following form

$$\begin{cases} \forall i \in [n], & \sum_{j\in[n']} c_{i,j}\ u_i^{In}\ u_j^{Out} = \ t_i^{In}; \\ \forall j \in [n'], & \sum_{i\in[n]} c_{i,j}\ u_i^{In}\ u_j^{Out} = \ t_j^{Out}, \end{cases}$$

where we have split the variable $\boldsymbol{u}$ in two vectors $\boldsymbol{u^{In}}$ and $\boldsymbol{u^{Out}}$, as corresponding to the constraints on the incoming traffic $\boldsymbol{t^{In}}$ and the outgoing traffic $\boldsymbol{t^{Out}}$, respectively. The solution of this system is a fixed point of the operator $T : (\boldsymbol{u^{In}}, \boldsymbol{u^{Out}}) \longrightarrow (\boldsymbol{v^{In}}, \boldsymbol{v^{Out}})$, where

$$\forall i \in [n],\ v_i^{In} := \frac{t_i^{In}}{\sum_j c_{i,j}\ u_j^{Out}} \qquad \text{and} \qquad \forall j \in [n'],\ v_j^{Out} := \frac{t_j^{Out}}{\sum_i c_{i,j}\ u_i^{In}}.$$

The reader can easily verify that the fixed point iterations of the operator T correspond exactly to the iterations of the dual matrix scaling (Algorithm 9.4.2), when the prior matrix is $H := [c_{i,j}]$, the row sums are given by $\overline{\boldsymbol{r}} := \boldsymbol{t^{In}}$ and the column sums by $\overline{\boldsymbol{c}} := \boldsymbol{t^{Out}}$. The variables $\boldsymbol{u}$ and $\boldsymbol{v}$ of the Brualdi iterations correspond respectively to $\boldsymbol{u^{In}}$ and $\boldsymbol{u^{Out}}$.

In fact, a straightforward generalization of this algorithm to an arbitrary network is possible when A has only $0/1$ entries. The polynomial equations $P(\boldsymbol{u}) = \boldsymbol{y}$ (9.8) take in that case the form (cf. Example 9.2.1):

$$\forall e \in [q], \quad \sum_{p \ni e} c_p \prod_{e' \in p} u_{e'} = y_e. \tag{9.11}$$

For all e, and for every OD pair p which is measured on link e ($p \ni e$), the product $\prod_{e'\in p} u_{e'}$ contains the factor u_e. As done previously for the star-shaped network, we can thus write the solution of the polynomial system (9.11) as a fixed-point of the operator T which maps

$(\mathbb{R}_+^*)^q$ onto itself, and is defined by:

$$\forall e \in [q], \quad T_e(\boldsymbol{u}) = \frac{y_e}{\sum_{p \ni e} c_p \prod_{\substack{e' \in p \\ e' \neq e}} u_{e'}}. \tag{9.12}$$

The polynomial equations (9.11) are equivalent to the fixed point equations : $\boldsymbol{u} = T(\boldsymbol{u})$. In what follows, we also denote by Q the denominator of T:

$$Q_e(\boldsymbol{u}) := \sum_{p \ni e} c_p \prod_{\substack{e' \in p \\ e' \neq e}} u_{e'}. \tag{9.13}$$

If T is nonexpansive, algorithms such as

$$\boldsymbol{u_{n+1}} = T(\boldsymbol{u_n})$$

are likely to converge. But the following proposition makes it difficult for T to be nonexpansive. First of all, let us introduce the partial Thomson metric as well as some other definitions [AGLN06]:

Definition 9.5.1. The partial *Thomson* metric d_T is defined on $(\mathbb{R}_+^*)^q$ as :

$$d_T(\boldsymbol{x}, \boldsymbol{y}) = \log(\max_{i \in [q]}(\frac{x_i}{y_i}, \frac{y_i}{x_i})).$$

Definition 9.5.2. An application f which maps $(\mathbb{R}_+^*)^q$ onto itself is said to be d_T-*nonexpansive* when

$$\forall \boldsymbol{x}, \boldsymbol{y} \in (\mathbb{R}_+^*)^q, \ d_T(f(\boldsymbol{x}), f(\boldsymbol{y})) \leq d_T(\boldsymbol{x}, \boldsymbol{y}).$$

Definition 9.5.3. An application $f : (\mathbb{R}_+)^q \longrightarrow (\mathbb{R}_+)^q$ is said to be *decreasingly subhomogeneous* when

$$\forall \lambda \geq 1, \forall \boldsymbol{x} \in (\mathbb{R}_+)^q, \ f(\lambda \boldsymbol{x}) \geq \lambda^{-1} f(\boldsymbol{x}),$$

where the latter inequality is component-wise.

Definition 9.5.4. An application $f : (\mathbb{R}_+)^q \longrightarrow (\mathbb{R}_+)^q$ is said to be *increasing* (resp. *decreasing*) when

$$(\boldsymbol{x} \leq \boldsymbol{y}) \Longrightarrow (f(\boldsymbol{x}) \leq f(\boldsymbol{y})) \quad (\text{resp. } f(\boldsymbol{x}) \geq f(\boldsymbol{y})),$$

where the inequalities are component-wise.

We still need a few lemmas to prove the next proposition :

Lemma 9.5.5. *Assume that $f : (\mathbb{R}_+^*)^q \longrightarrow (\mathbb{R}_+^*)^q$ is* decreasing. *Then, f is decreasingly subhomogeneous if and only if f is d_T-nonexpansive.*

Proof. Let $f : (\mathbb{R}_+^*)^q \longrightarrow (\mathbb{R}_+^*)^q$ be a decreasing application.

We first assume that f is decreasingly subhomogeneous (DSH). let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two vectors in $(\mathbb{R}_+^*)^q$, and λ be a real with $\lambda \geq \max_i(\frac{x_i}{y_i}, \frac{y_i}{x_i})$, so that $\log(\lambda) \geq d_T(\boldsymbol{x}, \boldsymbol{y})$.
We have :

$$\begin{cases} \boldsymbol{y} \leq \lambda \boldsymbol{x} \\ \boldsymbol{x} \leq \lambda \boldsymbol{y}. \end{cases}$$

Hence, $\boldsymbol{y} \leq \lambda^2 \boldsymbol{y}$ and $\lambda \geq 1$. Thus, we can use the DSH assumption for f :

$$\begin{aligned} & \begin{cases} f(\boldsymbol{y}) \geq f(\lambda \boldsymbol{x}) \geq \lambda^{-1} f(\boldsymbol{x}) \\ f(\boldsymbol{x}) \geq f(\lambda \boldsymbol{y}) \geq \lambda^{-1} f(\boldsymbol{y}) \end{cases} \\ \Longrightarrow \quad & \begin{cases} \forall i, \frac{f_i(\boldsymbol{x})}{f_i(\boldsymbol{y})} \leq \lambda \\ \forall i, \frac{f_i(\boldsymbol{y})}{f_i(\boldsymbol{x})} \leq \lambda \end{cases} \\ \Longrightarrow \quad & \max_i \left(\frac{f_i(\boldsymbol{x})}{f_i(\boldsymbol{y})}, \frac{f_i(\boldsymbol{y})}{f_i(\boldsymbol{x})} \right) \leq \lambda \\ \Longrightarrow \quad & \log(\lambda) \geq d_T(f(\boldsymbol{x}), f(\boldsymbol{y})) \end{aligned}$$

When $\log(\lambda) \longrightarrow d_T(\boldsymbol{x}, \boldsymbol{y})$, we obtain : $d_T(f(\boldsymbol{x}), f(\boldsymbol{y})) \leq d_T(\boldsymbol{x}, \boldsymbol{y})$

Conversely, assume that f is d_T-nonexpansive. Let $\boldsymbol{x} \in (\mathbb{R}_+)^q$ and $\lambda \geq 1$. Let $\boldsymbol{y} = \lambda \boldsymbol{x}$, such that $d_T(\boldsymbol{x}, \boldsymbol{y}) = \log(\lambda)$. We have $\boldsymbol{y} \geq \boldsymbol{x}$, thus $f(\boldsymbol{y}) \leq f(\boldsymbol{x})$ (f is decreasing). We now use the d_T-nonexpansiveness of f:

$$\log(\lambda) = d_T(\boldsymbol{x}, \boldsymbol{y}) \geq d_T(f(\boldsymbol{x}), f(\boldsymbol{y})) = \log \left(\max_i \frac{f_i(X)}{f_i(Y)} \right),$$

from which we deduce:

$$\begin{aligned} & \forall i \in [q], \ \lambda \geq \frac{f_i(\boldsymbol{x})}{f_i(\boldsymbol{y})} \\ \Longrightarrow \quad & \lambda f(\boldsymbol{y}) \geq f(\boldsymbol{x}) \\ \Longrightarrow \quad & f(\lambda \boldsymbol{x}) \geq \lambda^{-1} f(\boldsymbol{x}). \end{aligned}$$

□

Lemma 9.5.6. *T is decreasing. Moreover, T is decreasingly subhomogeneous if and only if its denominator Q is subhomogeneous, i.e.*

$$\forall \lambda \geq 1, \forall \boldsymbol{u} \in (\mathbb{R}_+^*)^q, \ Q(\lambda \boldsymbol{u}) \leq \lambda Q(\boldsymbol{u}).$$

Proof. This is trivial from the definition of T (9.12). □

Proposition 9.5.7. *T is d_T-nonexpansive if and only if every OD pair $p \in [m]$ is of length at most 2.*

Proof. Thanks to the previous lemmas, we only need to show that

$$(\forall p \in [m],\ \text{length}(p) < 2) \iff (Q \text{ is subhomogeneous}).$$

If every OD is of length 2 or less, then we can rewrite Q (see Expression (9.13)) as

$$\forall e \in [q],\ Q_e(\boldsymbol{u}) = c_{(e)} + \sum_{p=\{e,e'\}} c_p u_{e'}$$

where $c_{(e)}$ is the prior traffic on the OD which traverses only the edge e, (we set $c_{(e)}$ to 0 if no such OD exists), and $p = \{e, e'\}$ represents the OD pair comprising the two links e and e'. Thus, for $\lambda \geq 1$, we have

$$\lambda Q_e(\boldsymbol{u}) - Q_e(\lambda \boldsymbol{u}) = \lambda c_{(e)} - c_{(e)} \geq 0.$$

Therefore, Q is subhomogeneous. Conversely, assume that there are some ODs of length larger than 2. Let $n(e)$ be the length of the longest road that traverses e, and $\mathcal{P}_e^k$ be the set of all pairs of length k which traverse e. We can rewrite Q as:

$$Q_e(\boldsymbol{u}) = \sum_{k=1}^{n(e)} \sum_{\substack{p=\{e,e'_1,\dots,e'_{k-1}\} \\ p \in \mathcal{R}_e^k}} c_p u_{e'_1} \cdots u_{e'_{k-1}}.$$

Thus, for $\lambda \geq 1$ and $n(e) > 2$, we have

$$\lambda Q_e(\boldsymbol{u}) - Q_e(\lambda \boldsymbol{u}) = \sum_{k=1}^{n(e)} (\lambda - \lambda^{k-1}) \sum_{\substack{p=\{e,e'_1,\dots,e'_{k-1}\} \\ p \in \mathcal{P}_e^k}} c_p u_{e'_1} \cdots u_{e'_{k-1}} = O_{\lambda \to \infty}(-\lambda^{n(e)-1})$$

Hence, if λ is large enough, $\lambda Q(\boldsymbol{u}) < Q(\lambda \boldsymbol{u})$, and Q is not subhomogeneous. □

Note that in the star-shaped Network, every OD has length 2, which explains why Brualdi iterations converge. For arbitrary networks however, the condition on the length of the OD is very restrictive. Hence, we should rather use other kind of generalizations of the matrix balancing algorithm, which we next present.

9.5.2 Bregman's Balancing Method

A large class of Algorithms to solve unconstrained minimization problems is called the coordinate-descent: At each step of the computation, the objective function is minimized along *one coordinate only*, and this is repeated cyclically for each coordinates of the variable. In our case, when strict convexity of g is achieved (Assumption $(A.3)$), minimizing the objective function $g(\boldsymbol{\lambda})$ for the coordinate λ_i is equivalent to solving in λ_i the equation $\frac{\partial g}{\partial \lambda_i}(\boldsymbol{\lambda}) = S_i(\boldsymbol{\lambda}) - y_i = 0$, when every other λ_j $(j \neq i)$ is considered as constant. This method (Algorithm 9.5.1) was attributed to Bregman for its similarity with the dual form of

matrix balancing [Bre67a] (and this of the fixed-point algorithm studied in previous section, cf. Remark 9.5.1).

Algorithm 9.5.1 Bregman's Balancing Method

Choose $\boldsymbol{\lambda}^{(\mathbf{0})} \in \mathbb{R}^q$ and $\epsilon > 0$ sufficiently small
$t \leftarrow 0$
repeat
  $i \leftarrow (t \mod q) + 1$
  Find the unique solution μ of the following equation:

$$S(\lambda_1^{(t)}, \ldots, \lambda_{i-1}^{(t)}, \mu, \lambda_{i+1}^{(t)}, \ldots, \lambda_q^{(t)}) = y_i.$$

  $\boldsymbol{\lambda}^{(\boldsymbol{t+1})} \leftarrow \boldsymbol{\lambda}^{(\boldsymbol{t})}$
  $\lambda_i^{(t+1)} \leftarrow \mu$
  $t \leftarrow t + 1$
until $\|\nabla g(\boldsymbol{\lambda}^{(\boldsymbol{t})})\| < \epsilon$
Stop with the $\epsilon-$optimal Lagrange multiplier $\boldsymbol{\lambda}^* = \lambda^{(t)}$ and obtain the primal solution $\boldsymbol{x}^*$ according to Equation (9.5).

convergence It is shown in [AO82] that descent coordinates methods have a linear rate of convergence as soon as the objective function is strictly convex. This is guaranteed by Assumption $(A.3)$.

Remark 9.5.1. When the matrix A has only $0/1$ elements, we can give the solution of the equation in Algorithm 9.5.1 in closed-form: the equation becomes

$$\sum_{p \ni e} c_p \exp(\sum_{e' \in p} \lambda_{e'}^{(t)}) = y_e,$$

where the unknown is $\lambda_e^{(t)}$, and we have:

$$\lambda_e^{(t)} = \log \left(\frac{y_e}{\sum_{p \ni e} c_p \ \exp(\sum_{e' \in p,\ e' \neq e} \lambda_{e'}^{(t)})} \right)$$

If we use the notation $\boldsymbol{u} = \exp(\boldsymbol{\lambda})$, this gives the update rule:

$$\boldsymbol{u}_e^{(t+1)} = \frac{y_e}{\sum_{p \ni e} c_p \prod_{\substack{e' \in r \\ e' \neq e}} u_{e'}^{(t)}}$$

The similarity with the fixed-point iterations of the operator T studied in Section 9.5.1 is striking. The only change is the update pattern : unlike fixed points iterations, where every coordinate of $\boldsymbol{u}$ were updated as the same time, Bregman's iteration consist in a *Gauss-Siedl* pattern with the same operator T, that is to say that the coordinates are updated one at a time.

9.5.3 Iterative proportional Fitting

Unlike the previous method, the Iterative Proportional Fitting (IPF) makes no use of the dual form of the problem, and is probably the one that has been the most used in the area of traffic matrices estimation [CDVY00, LTY06, ZRDG03].

We denote by L_i the (truncated) hyperplane of all vectors $\boldsymbol{x} \geq 0$ verifying the i^{th} row of the system $\boldsymbol{y} = A\boldsymbol{x}$, that is to say

$$L_i = \{\boldsymbol{x} \geq \mathbf{0} : \ \boldsymbol{a_i}^T \boldsymbol{x} = y_i\},$$

where $\boldsymbol{a_i}^T$ is the i^{th} row of the matrix A, and by $\mathcal{V}$ the intersection of those hyperplanes: $\mathcal{V} = \bigcap_{i \in [q]} L_i$. The problem (9.3) becomes

$$\begin{array}{ll} \min & D(\boldsymbol{x} \| \boldsymbol{c}) \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{V} = \bigcap_{i \in [q]} L_i \end{array}$$

There's a wide literature [Bre67b, BBL95, BB96, DH94] about methods called cyclic projections, used to compute the (Kullback-Leibler) projection of a vector onto the intersection of several hyperplanes. In these methods, we compute at each iteration projection $\boldsymbol{x^{(t+1)}}$ of the current variable $\boldsymbol{x^{(t)}}$ onto the hyperplane L_i, where i is an index that goes cyclically through $[q]$

To find the projection $\hat{\boldsymbol{x}}$ of $\boldsymbol{x}$ onto L_i, the reader can verify that there must exist a scalar Lagrange multiplier $\boldsymbol{\lambda}$ such that

$$\forall p \in [m], \hat{x}_p = c_p \ \exp(\lambda A_{i,p}),$$

and λ can be computed by substituting the latter expression in $\boldsymbol{a_i}^T \boldsymbol{x} = y_i$. :

$$\sum_{p \in [m]} A_{i,p} c_p \ \exp(\lambda A_{i,p}) = y_i \tag{9.14}$$

If the matrix A has only $0/1-$elements, the reader can verify that the Kullback Leibler projection $\hat{\boldsymbol{x}}$ of $\boldsymbol{x}$ on L_e is given by :

$$\forall p \in [m], \ \hat{x}_p = \begin{cases} x_p \ \frac{y_e}{A_e \boldsymbol{x}} & \text{if } p \ni e; \\ x_p & \text{otherwise.} \end{cases} \tag{9.15}$$

There is a variant of this algorithm, called MART (Multiplicative Algebraic Reconstruction Technique) which generalizes the update given above for observation matrices whose

all elements are in the interval $[0, 1]$. This method was introduced by Gordon, Bender and Herman [GBH70] for an application to image reconstruction. The idea is to take a first-order approximation of Equation (9.14), from which we obtain:

$$\hat{x}_p = x_p \left(\frac{y_i}{A_i \boldsymbol{x}}\right)^{A_{i,p}}$$

Algorithm 9.5.2 Iterative Proportional Fitting (or MART)

Choose $\epsilon > 0$ sufficiently small
$t \longleftarrow 0$
$\boldsymbol{x}^{(\mathbf{0})} \longleftarrow \boldsymbol{c}$
repeat
  $i \longleftarrow (t \ mod \ q) + 1$
  for $p \in [m]$ **do**
    $x_p^{(t+1)} \longleftarrow x_p^{(t)} \left(\frac{y_i}{A_i \boldsymbol{x}^{(t)}}\right)^{A_{i,p}}$
  end for
  $t \longleftarrow t + 1$
until $\|(A\boldsymbol{x}^{(t)} - \boldsymbol{y})\| < \epsilon$
Stop and set $\boldsymbol{x}^* = \boldsymbol{x}^{(t)}$.

rate of convergence It was shown [BBL95] that cyclic projection methods (for Euclidean projections) do converge at a linear convergence rate given by the "angle" between the hyperplanes. Similarly, Iusem [Ius91] worked on the convergence of cyclic projections with the Kullback-Leibler divergence and proved a linear convergence. This time, the rate depends on a geometric parameter θ given by

$$\theta = \inf_{\boldsymbol{x} \notin \mathcal{V}} \frac{\max_{i \in [q]} d_Q(\boldsymbol{x}, L_i)}{d_Q(\boldsymbol{x}, \mathcal{V})} \tag{9.16}$$

In this expression, $d_Q(\cdot, \cdot)$ is the distance associated to the scalar product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{x}^T Q \boldsymbol{y}$, where Q is the hessian of the cross-entropy, calculated at the optimal point $\boldsymbol{x}^*$ of Problem (9.3), i.e. $Q = \mathrm{Diag}(1/x_1^*, ..., 1/x_m^*)$. Iusem showed [Ius91] that $\theta \in [0, 1]$ and the sequence $(\boldsymbol{x}^{(t)})_{t \in \mathbb{N}}$ generated by the cyclic projection algorithm (i.e. Algorithm 9.5.2 when A is $0/1$) converges with a rate no worse than $\rho = \frac{q}{q+\theta^2}$ (for the norm d_Q). Elfving [Elf80] proved that the convergence of the MART algorithm is also linear for observation matrices A with fractional components.

Remark 9.5.2. When the matrix A has only $0/1$ entries, the algorithms 9.5.1 and 9.5.2 generate the same sequence of variable $\boldsymbol{x}^{(t)}$. The connection between these algorithms was established by Censor et. al. [CDPE+90] (see also [FRT97]). In fact, Bregman's balancing method 9.5.1 is to the IPF algorithm 9.5.2 as the dual method for matrix balancing 9.4.2 is to the primal 9.4.1. The main difference between the former (matrix balancing algorithms) and the latter algorithms (entropy projections) is the order in which we update the variables (simultaneously or cyclically).

9.6 Second order methods

In order to achieve a better performance, we could use Newton iterations to solve the unconstrained optimization problem (9.6). It is well known that Newton algorithm exhibits a local quadratic convergence, but the global convergence is not always achieved. We faced a lot of issues, even on small examples, when the initial guess $\boldsymbol{\lambda}^{(0)}$ is far away from the optimal solution $\boldsymbol{\lambda}^*$ and the Newton sequence diverges.

Algorithm 9.6.1 Curved-Search Descent

$t \leftarrow 0$
$\boldsymbol{\lambda^0} = \mathbf{0} \in \mathbb{R}^q$
repeat
  Choose some parameters $\alpha^{(t)}$ and $\beta^{(t)}$
  $\boldsymbol{h^{(t)}} \leftarrow \nabla g(\boldsymbol{\lambda^{(t)}})$
  $H^{(t)} \leftarrow \nabla^2 g(\boldsymbol{\lambda^{(t)}})$
  $\boldsymbol{d^{(t)}} \leftarrow -\beta^{(t)} \frac{\|\boldsymbol{h^{(t)}}\|^2 [H^{(t)}]^{-1}}{(\boldsymbol{h^{(t)}})^T \ [H^{(t)}]^{-1} \ \boldsymbol{h^{(t)}}} \boldsymbol{h^{(t)}}$
  $\boldsymbol{z^{(t)}} \leftarrow -\alpha^{(t)} \|\boldsymbol{h^{(t)}}\| \boldsymbol{h^{(t)}}$
  Set $\boldsymbol{\lambda^{(t+1)}}$ as the minimizer of g along the quadratic curve $u \mapsto \boldsymbol{\lambda^{(t)}} + u\boldsymbol{d^{(t)}} + \frac{u^2}{2}\boldsymbol{z^{(t)}}$
  $t \leftarrow t + 1$
until $\|\boldsymbol{h^{(t)}}\| < \epsilon$
Stop with the $\epsilon-$optimal Lagrange multiplier $\boldsymbol{\lambda^*} = \lambda^{(t)}$ and obtain the primal solution X^* according to Equation (9.5).

For this reason,as proposed in [FRT97], we can use a curved-search algorithm, which is a mix between Newton's method and the gradient descent method. Curved-Search algorithms have been introduced by Ben-Tal, Melman and Zowe [BTMZ90] in order to solve unconstrained convex minimization programs with a quadratic rate of convergence. In fact, as explained in [FRT97] this algorithm turns out to be a nonlinear combination of the *signed* Newton's method and the steepest descent direction: At each step of Algorithm 9.6.1, $\boldsymbol{d^{(t)}}$ is a Newton's direction, and $\boldsymbol{z^{(t)}}$ is a steepest gradient direction. The difficulty in this algorithm is to determine appropriate values for the nonnegative parameters $\alpha^{(t)}$ and $\beta^{(t)}$. One should notice that when $\alpha^{(t)} = 0$, the iteration described in Algorithm 9.6.1 is simply a gradient step, and when $\beta^{(t)} = 0$, this is a Newton step. In [BTMZ90], the authors study a version of the latter algorithm which converges to the unique minimum of the strictly convex function g at a quadratic rate.

Chapter 10

Optimization of Netflow measurements

We address in this chapter the problem of optimizing the use of Network monitoring tools, such as Netflow (Cisco Systems), on a large IP network, and we present in greater details the results of [SBG10, SGB10]. Some results of this chapter, including our *experimental design* formulation of the optimal monitoring problem, were presented at the conference [BGS08].

We shall see that the theory of optimal experimental design, studied in Part I of this manuscript, is a natural framework for both the combinatorial problem of selecting the "best" subset of interfaces on which Netflow should be activated, and the problem of finding the optimal sampling rates of the network-monitoring tool on these interfaces. The main issue is the size of the matrices involved in this problem, which are of size $n^2 \times n^2$ on a network with n nodes. Both SDP and multiplicative algorithms approaches fail to be efficient, as seen in Chapter 6.

We develop a new method, which reduces to solving a stochastic sequence of Second Order Cone Programs. From a theoretical point of view, our approach is actually equivalent to compute an experimental design for a new design criterion, which is defined as the expected value of the $\boldsymbol{c}-$optimal designs, when the vector $\boldsymbol{c}$ is drawn from a Gaussian distribution. We approximate this design by taking the mean of several $\boldsymbol{c}-$optimal designs, a scheme which we have called "Successive $\boldsymbol{c}-$optimal designs" (SCOD). The motivation for this new method resides mainly in the size of the problems which it can handle: we have seen in Chapter 6 that it is possible to solve very large instances of $\boldsymbol{c}-$optimal design problems by SOCP. Interestingly, there are also some heuristic arguments which let us think that the SCOD approximates the classic $A-$optimal design. We will show by examples that this fact is verified in practice, and we will derive some bounds between the SCOD and the $A-$optimal design in simple cases.

We next give experimental results relying on real data from both Abilene and the Open-transit network of France Telecom, which show that our approach can be used for instances that were previously intractable, and we compare our method to previously proposed ones:

- Several networks are not (or only partially) instrumented with routers that support Netflow. If an Internet provider wishes to equip a number of additional routers with Netflow, an interesting problem is thus to identify the most meaningful subset of locations for the monitoring-tool. We shall compare our SCOD approach to the greedy algorithm [SQZ06] for this problem.
- We next evaluate our approach for the problem of selecting the optimal sampling rates of Netflow. The Internet provider typically sets a threshold the number of packets that may be sampled at each router location during a given period of time. The goal is thus to allocate optimal sampling rates to the incoming interfaces of each router, while keeping the number of sampled packets under the threshold. For the Opentransit network (with $m = 13456$ OD pairs), we do not know any other algorithm which can handle this optimization problem.
- In a dynamic context, sampling rates should be optimized by taking into accounts the errors on the past measurements. To this end, Singhal and Michailidis [SM08] proposed to introduce in the information matrix of every design an additional term which accounts for the covariance on the past measurements, and is computed via a Kalman filter. We will show how to adapt this method to larger networks, thanks to our SCOD method. In fact, we shall see by an example on the Abilene network that in situations where the traffic has a very high variability, it is better to ignore the impact of past measurements.

10.1 Background

10.1.1 Netflow measurements

We have seen in Chapter 8 that the problem of estimating the traffic matrix from link counts is ill-posed, and we require additional information to solve this problem. A way to introduce new constraints is to use a network-monitoring tool such as Netflow (Cisco Systems), cf. Section 8.3.1. Of course, activating Netflow everywhere on the network yields an extensive knowledge of the OD flows. According to [CIS07] however, activating Netflow on an interface of a router causes its CPU load to increase by 10 to 50%. It is now possible to use a sampled version of Netflow, which substantially decreases both the CPU utilization and the bandwidth consumption caused by Netflow. It was shown indeed [CB05] that the overhead involved by Netflow is roughly proportional to the sampling rates. The counterpart of the sampling is of course that sampled measurements yield less accurate estimations of the traffic. It is thus of great interest to optimize the use of this tool. The problem is both to decide where to activate Netflow, and at which sampling rate.

Most operators collect Netflow information for multiple purposes, such as security or billing, not only for estimating the traffic. However, we believe that the present approach, which addresses the latter goal, might also be of some interest for other purposes, since it indicates which routers or interfaces captures the most valuable information about the

traffic. Moreover, we will see that this approach leads to a nice mathematical formulation and to scalable algorithms.

Recall that when Netflow is activated on an interface of the network, it analyzes the headers of the packets traversing this interface and collects some statistics, such as the source and destination IP addresses of these packets (cf. Section 8.3.1). However, we are not trying to infer the global path of the packets from IP source to IP destination, but only the part of their path which is inside the network of interest, like the backbone of an autonomous system (AS). In the sequel, we will use the terms *internal source* and *internal destination* to refer to the ingress and egress routers of a packet within the backbone of interest.

Practically, we will assume throughout this paper that when Netflow performs a measurement on the k^{th} interface, we are able to break out the flows traversing this interface according to their internal destination. This results in a multidimensional observation $\boldsymbol{y_k}$, whose entry d is the sum of all the flows traversing k and having the destination d. The model is linear (cf. Example 10.1.1):

$$\boldsymbol{y_k} = A_k \boldsymbol{x} \ . \tag{10.1}$$

Note that this assumption is more general and more realistic than the one made

Example 10.1.1. We observe Netflow records on the link Houston → LA of the Abilene backbone. This link is used by the flows Kansas City→LA, Houston→Seattle, and by all the flows from any one of Houston, Atlanta, or Washington to either LA or Sunnyvale.

Since Netflow "breaks" the flows with respect to their destination, we obtain three partial sums of the flows listed above (in which the ODs are grouped with respect to their internal destination):

$$\boldsymbol{y}^{\text{Houston}}_{\to \text{LA}} = \begin{pmatrix} \theta^{\text{Houston}}_{\to \text{LA}} & +\theta^{\text{Atlanta}}_{\to \text{LA}} & +\theta^{\text{Washington}}_{\to \text{LA}} & +\theta^{\text{Kansas}}_{\to \text{LA}} \\ \theta^{\text{Houston}}_{\to \text{Sunnyvale}} & +\theta^{\text{Atlanta}}_{\to \text{Sunnyvale}} & +\theta^{\text{Washington}}_{\to \text{Sunnyvale}} & \\ \theta^{\text{Houston}}_{\to \text{Seattle}} & & & \end{pmatrix}.$$

These three partial sums obviously carry more information on the traffic matrix than the SNMP data corresponding to the same link (which consists in the sum of *all* the flows). For example, Netflow measurements yield a direct estimation of the traffic from Houston to Seattle (third entry of $\boldsymbol{y}^{\text{Houston}}_{\to \text{LA}}$).

in [SM08], where the authors assume that when the monitoring tool analyzes a packet, it is able to find both its internal source and destination. In practice, one can find the internal destination of a packet by simulating the path toward its ultimate destination with the forwarding tables of the routers, but finding the internal source of a packet is a challenging issue. The main difference with the simplified model considered in [SM08] is that the information matrices $A_k^T A_k$ are not diagonal anymore, which makes the problem much harder computationally.

A precise description of the classical methodology used to infer the origin-destination traffic from Netflow measurements is made in [FGL+01]. It is common to activate Netflow only on ingress links of a backbone in order to cope with the uncertainty on the internal source of the packets. In this paper, we show that other deployment strategies can be useful.

10.1.2 Related work

Many authors from the network research community investigated the placement of Netflow. We briefly review their contributions:

Zang and Nucci [ZN05] posed the Netflow placement problem as an Integer program whose objective is to minimize the cost of deployment of the monitoring tool on the network, taking into account the costs required to upgrade the routers so that they support Netflow. In this approach, the constraint imposes that Netflow monitors at least a fraction α of all the traffic. They proposed two greedy heuristics in order to find a near-optimal solution to this NP-Hard integer program.

Bouhtou and Klopfenstein [BK07] pursued this approach by taking into account the variations of the traffic in time. In most networks, the routing table is not static indeed, and the placement of Netflow must be robust to possible routing modifications. To tackle this issue, the authors formulated an optimization problem with probability constraints which can be approximated by a sequence of integer linear programs.

Cantieni, Iannaccone, Barakat, Diot and Thiran [CIB+06] interested themselves in the optimal rates at which Netflow should be sampled on each router. They formulated a convex optimization problem, in which the probability that a packet is intercepted by Netflow is maximized. They solve this problem by a projected gradient algorithm.

Bermolen, Vaton and Juva [BVJ06] were the first to investigate the optimal placement of Netflow in light of the *experimental design* background. Based on the model proposed by Cao et al. [CDVY00], they suggested that the observation vector $\boldsymbol{y}$ has a normal distribution, whose expected value and covariance matrix depends on the expected value $\overline{\boldsymbol{x}}$ of the OD flows, and derived the Fisher information matrix for any placement of the measures. The authors of [BVJ06] give a scheme for selecting a few interfaces on which Netflow should be activated in priority.

Song, Qiu and Zhang [SQZ06] used classical criteria from the theory of experimental design to choose a subset of interfaces where Netflow should be activated, and developed

an efficient greedy algorithm to find a near optimal solution to this combinatorial problem. In Chapter 7, we have shown indeed that the greedy algorithm always finds a solution within $1 - 1/e \approx 62\%$ of the optimum.

Singhal and Michailidis [SM08] considered a state-space model representing the evolution of the traffic matrix over time, in which the estimation of the traffic can be done by a Kalman filter. They successfully applied the experimental design theory to formulate the problem of finding the sampling rates that minimize the covariance matrix of the Kalman filter as a Semidefinite Program (SDP). Since the covariance matrix is computed recursively in the filtering process, it contains information on the past measurements, and computing new sampling rates at each time step makes the estimation more and more accurate.

10.2 Experimental design formulation of the problem

10.2.1 Netflow optimal deployment

Let $\mathcal{I} = \{1, \dots, s\}$ be the set of all interfaces on which Netflow can be activated. We start with the discrete problem, in which the operator wants to choose a subset of these interfaces for the Netflow measurements. Note that this problem is also meaningful when a network is not yet or is only partially instrumented with routers supporting Netflow, and when the Internet provider wants to equip a number of additional routers with a network-monitoring tool.

We denote by $\mathcal{I}^{\mathrm{a}}$ the set of interfaces on which Netflow is activated. The measurement vector $\boldsymbol{y}$ is now the concatenation of the SNMP data $\boldsymbol{y}^{\mathrm{SNMP}}$ with all the Netflow measurements $(\boldsymbol{y}_k)_{k \in \mathcal{I}^{\mathrm{a}}}$. We define the *design* variable $\boldsymbol{w}$ as the $0/1$ vector of size s, where w_k equals 1 if and only if $k \in \mathcal{I}^{\mathrm{a}}$. The measurements are never exact in practice, and so we have to deal with a noise ϵ, which is a result, among other things, of lost packets, misalignment of SNMP polling intervals, and Netflow sampling. This can be modeled as follows:

$$\boldsymbol{y} = A_{\boldsymbol{w}}\, \boldsymbol{x} + \boldsymbol{\epsilon}, \tag{10.2}$$

$$\text{where } \boldsymbol{y} = \begin{pmatrix} \boldsymbol{y_0} \\ \boldsymbol{y_{k_1}} \\ \vdots \\ \boldsymbol{y_{k_n}} \end{pmatrix} \text{ and } A_{\boldsymbol{w}} := \begin{bmatrix} A_0 \\ A_{k_1} \\ \vdots \\ A_{k_n} \end{bmatrix}.$$

In the latter observation equation, we have used the index 0 to refer to the SNMP measurements, which are available in any case ($\boldsymbol{y_0} = \boldsymbol{y}^{\mathrm{SNMP}}$ and $A_0 = A$). We assume here that the noises on the observations are mutually independent, that is to say that the covariance matrix $\Sigma = \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]$ is known and has only diagonal entries. To simplify the notation, we will assume that $\Sigma = \boldsymbol{I}$ (one may always reduce to this case with a left scaling (by $\Sigma^{-1/2}$) of $\boldsymbol{y}, A_{\boldsymbol{w}}$ and $\boldsymbol{\epsilon}$). Note that the vector of flow volumes $\boldsymbol{x}$ is the unknown in this

problem, and plays the role of the parameter $\boldsymbol{\theta}$ in the classic experimental design problems studied in the first part of this thesis.

This observation model is clearly of the same kind of that studied in Chapter 7 (cf. Equation (7.1)). Therefore, the experimental design approach consists in choosing the design vector $\boldsymbol{w}$ so as to *maximize* the information matrix of the design:

$$\begin{aligned} M(\boldsymbol{w}) = (\operatorname{Var} \hat{\boldsymbol{x}})^{-1} &= A_{\boldsymbol{w}}^T A_{\boldsymbol{w}} \\ &= A_0^T A_0 + \sum_{k=1}^{s} w_k A_k^T A_k, \end{aligned} \tag{10.3}$$

where $\hat{\boldsymbol{x}}$ is the best linear unbiased estimator of $\boldsymbol{x}$:

$$\hat{\boldsymbol{x}} = \left(A_{\boldsymbol{w}}^T A_{\boldsymbol{w}}\right)^{-1} A_{\boldsymbol{w}}^T \boldsymbol{y}. \tag{10.4}$$

We can now give a mathematical formulation to the problem of optimally deploying Netflow on no more than n interfaces:

$$\max_{\boldsymbol{w} \in \{0,1\}^s} \Phi\Big(M(\boldsymbol{w})\Big) \qquad \text{s.t.} \quad \sum_i w_i \leq n, \tag{10.5}$$

where Φ is any design criterion from the experimental design literature.

10.2.2 Optimal sampling rates

We now show that the optimal sampling problem can be formulated in the form of Problem (10.5), too. Following [SM08], we assume that Netflow performs a random sampling with rate w_k on the interface k (this is one of the possibilities to configure the sampling of Netflow, and considered as the best one in the *Netflow services solutions guide* [CISa]). As explained in Section 10.1.1, Netflow can be used to sort the packets with respect to their internal destination. Let N_{kd} be a counter that records the number of sampled packets from interface k which have the internal destination d. This number follows a binomial distribution with $(\boldsymbol{y_k})_d = A_{kd}\boldsymbol{x}$ trials and probability of success w_k, where A_{kd} is the row corresponding to the destination d in the matrix A_k. The best unbiased linear estimator of $\boldsymbol{y_k}$ is given by $(\hat{\boldsymbol{y}}_{\boldsymbol{k}})_d = w_k^{-1} N_{kd}$, and we have:

$$\begin{aligned} \operatorname{Var}(\hat{\boldsymbol{y}}_{\boldsymbol{k}})_{d,d} = w_k^{-2} \operatorname{var}(N_{kd}) &= w_k^{-2}\ w_k(1-w_k)\ y_{kd} \\ &\approx w_k^{-1}\ (A_k \boldsymbol{x})_d, \end{aligned} \tag{10.6}$$

where the latter approximation is valid in the (expected) case where the sampling rates are small. The aggregate observation matrix is now $\tilde{A} := [A^T, A_1^T, \ldots, A_s^T]^T$ and does not depend on the design $\boldsymbol{w}$ anymore (we assume that measurements are performed on all interfaces: $\boldsymbol{y} = \tilde{A}\boldsymbol{x} + \epsilon$). Instead, the vector of sampling rates $\boldsymbol{w}$ is involved in the covariance matrix of the noise $\boldsymbol{\epsilon}$. We model the variance of the noise on the SNMP data

as $\sigma^2 I$ for a small parameter σ, while the variance of Netflow measurements follows from Equation (10.6):

$$\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \Sigma(\boldsymbol{w}) = \begin{pmatrix} \sigma^2 I & & & \\ & \frac{\mathrm{Diag}(A_1\boldsymbol{x})}{w_1} & & \\ & & \ddots & \\ & & & \frac{\mathrm{Diag}(A_s\boldsymbol{x})}{w_s} \end{pmatrix}.$$

As in the discrete case, we can make explicit the best unbiased linear estimate of the flows (which is given by the Gauss-Markov theorem 2.2.1), as well as the information matrix of the sampling design $\boldsymbol{w}$:

$$\hat{\boldsymbol{x}} = \left(\tilde{A}\Sigma(\boldsymbol{w})^{-1}\tilde{A}\right)^{-1}\tilde{A}^T\Sigma(\boldsymbol{w})^{-1}\boldsymbol{y}. \tag{10.7}$$
$$M(\boldsymbol{w}) = \tilde{A}^T\Sigma(\boldsymbol{w})^{-1}\tilde{A}.$$

Finally, we define the normalized observation matrices $\overline{A_0} = \sigma^{-1}A_0$ and $\overline{A_i} = \mathrm{Diag}(A_i\boldsymbol{x})^{-1/2}A_i$, so that the information matrix can be written as

$$M(\boldsymbol{w}) = \overline{A_0}^T\ \overline{A_0} + \sum_{k=1}^{s} w_k \overline{A_k}^T\ \overline{A_k}\ . \tag{10.8}$$

Hence, the $\Phi-$maximization of $M(\boldsymbol{w})$ takes a similar form as Problem (10.5), with a continuous variable $\boldsymbol{w}$ that is subject to linear constraints which we shall described in Section 10.2.3.

It remains to cope with the fact that the normalized observation matrices $\overline{A_i}$ explicitly depend on the unknown $\boldsymbol{x}$. Similarly to what is done in [SM08], we use a prior estimate of $\boldsymbol{x}$ to compute an approximate version of the $\overline{A_i}$. In the numerical studies presented in Section 10.5, we track the OD flows over time in a network, and we use the previous estimate $\hat{\boldsymbol{x}}_{t-1}$ in place of $\boldsymbol{x}_t$. At $t = 1$, we can use a tomogravity estimate of $\boldsymbol{x}$, which is a classical prior in the traffic matrix estimation literature (cf. Chapter 9).

10.2.3 Constraints on the sampling rates

Since the version 9 of Netflow, it is possible to set different sampling rates for each interface of a router where Netflow is activated. The Internet provider typically sets a threshold on the volume of packets to be analyzed with Netflow at each router location, so as to limit the overhead. For a specific router $\mathcal{R}$, the number of sampled packets can be approximated by

$$\sum_{k\in\mathcal{I}_\mathcal{R}} w_k f_k,$$

where the sum is carried out over the incoming interfaces $\mathcal{I}_\mathcal{R}$ of router $\mathcal{R}$, and f_k is the total number of packets traversing interface k (cf. Figure 10.1). In practice, f_k can be estimated

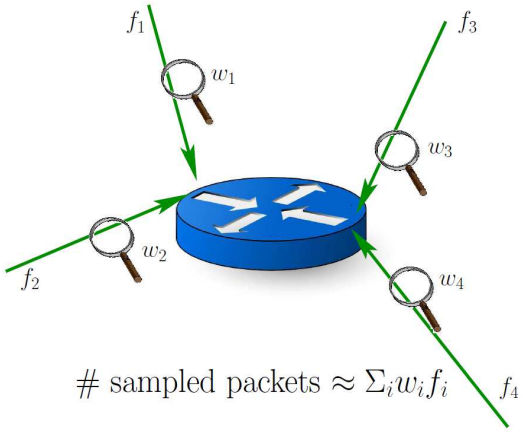


Figure 10.1: Per-router constraints. A target number of packets to be analyzed by Netflow is set by the ISP.

from previous values of the SNMP data. The constraints can thus be summarized as a set of linear inequalities of the form $R\boldsymbol{w} \leq \boldsymbol{d}$, where the entries of R depends on $\boldsymbol{y}^{\text{SNMP}}$ and $\boldsymbol{d}$ is a target set by the Internet provider. This is an alternative approach to that of Singhal and Michailidis [SM08], who use a matrix R depending only on the topology of the network.

10.3 Resolution of the problem: previous approaches

In this section, we review the previous methods that have been proposed to solve the discrete *Netflow optimal deployment* problem, as well as its continuous relaxation (*Netflow optimal sampling*). For simplicity of notation, we assume that the observation matrices have already been normalized by the left diagonal scaling mentioned in Section 10.2.1, so that $M(\boldsymbol{w})$ takes the form (10.3). This allows us to handle both problems in a unified framework. Note that any method which solves the continuous problem can be applied to obtain an approximate solution to the discrete problem, by applying simple rounding heuristics.

10.3.1 Greedy Algorithm

In the discrete case, and when there is a single constraint of the form $\sum_i w_i \leq n$, we can make use of a greedy algorithm, which is suggested by the results of [BGS08]. The principle is to start from $\mathcal{G}_0 = \emptyset$ and to construct sequentially the sets $\mathcal{G}_k := \mathcal{G}_{k-1} \cup \operatorname{argmax}_{i \in [s]} \varphi_p(\mathcal{G}_{k-1} \cup i)$, for $k = 1, ..., n$.

On a network with $m = 10^4$ OD pairs, the computation of the objective function $\Phi_p(\boldsymbol{w})$ requires about 5 minutes on a PC at 4GHz, since it involves the diagonalization of a $m \times m$ matrix. Consequently, selecting only one out of one hundred interfaces already requires

more than 3 hours. The authors of [SQZ06] proposed to use the special values $p = 0$ or $p = -1$ ($D-$ and $A-$optimal design), for which we can implement the Fedorov sequential design algorithm [Fed72], which computes efficiently the increment of the criterion thanks to Sherman-Morrison like formulae:

$$\begin{aligned}\Big(M + A_k^T A_k\Big)^{-1} &= M^{-1} - M^{-1}A_k^T\Big(I + A_k M^{-1} A_k\Big)^{-1} A_k M^{-1},\\ \det\Big(M + A_k^T A_k\Big) &= \det(M)\,\det\Big(I + A_k M^{-1} A_k^T\Big).\end{aligned}$$

At the beginning of the algorithm, the initial information matrix $M_0 = A_0^T A_0$ is not invertible. The authors of [SQZ06] remedy this problem by regularizing the initial observation matrix: they set $M_0 = A_0^T A_0 + \varepsilon I$, with $\varepsilon = 0.001$. Although this trick may look arbitrary, it leads to very good results.

If we leave aside the information from the SNMP measurements ($M_0 = \varepsilon I$), this algorithm performs astonishingly well, and the set of interfaces of a very large network can be ordered very quickly (it took 15 minutes on a PC at 4GHz to order in a greedy fashion the 116 routers of the Opentransit network with 13456 OD pairs). However, if we want to take into account the SNMP measurement (so as to avoid redundancy), M_0 is not sparse anymore, and small-rank updates become computationally expansive. The authors of [SQZ06] work on a similar experimental design problem, and store a sparse LU decomposition of M_k in place of the full matrix M_k^{-1}, which still allows one to compute $M^{-1}A_k^T$. In our case though, the LU decomposition is full and the greedy updates are intractable.

10.3.2 Semidefinite Programming

We have seen in Section 3.3 that the $E-$, $A-$ and $D-$ optimal design problems can be formulated as semidefinite programs (or as a MAXDET program). The great advantage of SDP approaches resides in the possibility to handle the resource constraints $R\boldsymbol{w} \leq \boldsymbol{d}$. This was noticed by Singhal and Michailidis [SM08], who have formulated the Netflow optimal sampling problem under per-router constraints as:

$$\begin{aligned}\min_{q} \quad & \sum_{j=1}^{m} q_j \\ s.t. \quad & \left(\begin{array}{c|c} M(\boldsymbol{w}) & \boldsymbol{e_j} \\ \hline \boldsymbol{e_j}^T & q_j \end{array}\right) \succeq 0, \quad j = 1, ..., m \\ & R\boldsymbol{w} \leq \boldsymbol{d},\ \boldsymbol{w} \geq 0,\end{aligned} \tag{10.9}$$

where $\boldsymbol{e_j}$ denotes the j^{th} vector of the canonical basis of $\mathbb{R}^m$ (the latter problem is for $A-$optimality). Singhal and Michailidis further proposed to add in $M(\boldsymbol{w})$ a constant term which accounts for the covariance matrix of the errors of the past measurements, and which is updated at each iteration by a Kalman filter.

However, this SDP is intractable by state-of-the-art solvers for networks with more than $m \approx 300$ OD pairs, as corresponding to $n = 17$ nodes. Therefore, we need a new, scalable method to solve the optimal sampling problem on large networks. We next present a new design criterion which can be approximated by a sequence of second order cone programs, even is the network is very large.

10.4 Successive $c-$Optimal Designs

The hardness of the optimal experimental design is linked to the large dimension of the parameter that we want to estimate, which leads to large size covariance matrices. Rather than estimating the full parameter $\boldsymbol{x}$, a natural idea is to estimate a linear combination $z = \boldsymbol{c}^T\boldsymbol{x}$ of the flows. This problem is called $\boldsymbol{c}-$optimal design (cf. Section 2.3.1), and consists in minimizing the (scalar) variance of the best linear unbiased estimator $\hat{z}$:

$$\mathrm{var}(\hat{z}) = \boldsymbol{c}^T M(\boldsymbol{w})^\dagger \boldsymbol{c},$$

where $M^\dagger$ denotes the Moore-Penrose inverse of M. Although scalar, the latter quantity still depends (non-linearly) on a $m \times m$ matrix. Hence, the semidefinite programming approach to solve this problem is intractable on large networks (cf. Chapter 6).

We have seen in Chapter 5 that the $\boldsymbol{c}-$optimal design problem actually reduces to a second order cone program (Theorem 5.2.3), which remains tractable on very large instances of the problem (cf. Chapter 6). Let us recall this result: The $\boldsymbol{c}-$optimal design problem (minimizing $\boldsymbol{c}^T M(\boldsymbol{w})^\dagger \boldsymbol{c}$ under the constraints $R\boldsymbol{w} \leq \boldsymbol{d}$) is equivalent to the following SOCP:

$$\min_{\boldsymbol{w},\ \boldsymbol{\mu},\ (\boldsymbol{h_i})_{i=0,\dots,s}} \quad \sum_{i=0}^{s} \mu_i \tag{10.10}$$

$$A_0^T \boldsymbol{h_0} + \sum_{i=1}^{s} A_i^T \boldsymbol{h_i} = \boldsymbol{c}$$

$$R\boldsymbol{w} \leq \boldsymbol{d},\ \boldsymbol{w} \geq \boldsymbol{0}$$

$$\left\| \begin{bmatrix} 2\boldsymbol{h_0} \\ 1 - \mu_0 \end{bmatrix} \right\| \leq 1 + \mu_0$$

$$\left\| \begin{bmatrix} 2\boldsymbol{h_i} \\ w_i - \mu_i \end{bmatrix} \right\| \leq w_i + \mu_i,\ (i = 1, \dots, s).$$

This theorem shows how to compute the optimal sampling rates $\boldsymbol{w}^*$ of the measurements (for the $\boldsymbol{c}-$combination of the flows) by SOCP. This can be done very efficiently with interior points codes such as SeDuMi [Stu99]. Moreover, this method takes advantage of the sparsity of the matrices A_i, while both the SDP approach and the multiplicative algorithms involve the information matrices $A_i^T A_i$, which are *not very sparse* in general.

In fact, Internet providers usually take advantage of Netflow measurements to estimate several OD flows (not only a linear combination of them). The approach which we present is a heuristic based on the computation of several $\boldsymbol{c}-$optimal designs (for example, for several vectors $\boldsymbol{c_i}$ drawn from a normal distribution). The motivation for this approach comes from this intuitive statement: if a design is good for the estimation of the linear combination of the flows $\boldsymbol{u}^T\boldsymbol{x}$ for every randomly generated vector $\boldsymbol{u}$ (from a normal distribution), then it should also be good for the estimation of $\boldsymbol{x}$. We describe our method in more details in the next section, and we give a heuristic argument for our approach in Section 10.4.2.

10.4.1 SCOD: a flexible scheme to select a design

Our method can be described by a parameter N, which indicates the number of $\boldsymbol{c}-$optimal designs to compute. Some vectors $\boldsymbol{c_1}, \dots, \boldsymbol{c_N} \in \mathbb{R}^m$ are selected by the experimenter. Then, a $\boldsymbol{c_i}-$optimal design $\boldsymbol{w}_{\boldsymbol{c_i}}$ is found by solving the SOCP (10.10) for each $i \in [N]$. Finally, we combine the resulting designs by taking the mean.

The parameter N can be adjusted by the experimenter: it should be large enough, so that the (generalized) Elfving set is *measured in several directions* (cf. Figure 5.1, page 92), and so that the average blur the particularities of each individual $\boldsymbol{c_i}-$optimal design, but it should remain small enough so as to keep the computation time reasonable.

An interesting feature of this method is the flexibility brought by the choice of the vectors $\boldsymbol{c_i}$: if the Internet provider attaches equal importance to each OD pair, a natural choice is to draw the vectors $\boldsymbol{c}$ from a normal distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$ (we call this scheme "*uniform SCOD*"). If only a subset of all the OD flows is of interest (e.g. the ODs #1, #19, and #31), then a possibility is to draw vectors $\boldsymbol{c_i}$ which have nonzero components only on the corresponding coordinates (e.g. $\boldsymbol{c} = K\boldsymbol{u}$, for $K = [\boldsymbol{e_1}, \boldsymbol{e_{19}}, \boldsymbol{e_{31}}]$ and $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_3)$).

Another possibility is to weight the importance of the different flows. For example, if σ_i reflects the importance that the Internet provider attaches to the accuracy of the estimation of the traffic on the i^{th} OD pair, then we can draw the vectors $\boldsymbol{c_i}$ with respect to $\mathcal{N}(\boldsymbol{0}, \mathrm{Diag}(\boldsymbol{\sigma}^2))$. In practice, the quality of an estimator $\hat{\boldsymbol{x}}$ for the traffic matrix is measured by the $L2$ error of the flows:

$$\mathrm{Rel}_2(\hat{\boldsymbol{x}}) = \frac{\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}, \tag{10.11}$$

which gives more weight to the accuracy of *heavy* flows. In Section 10.5.4, we therefore study the impact of generating the vectors $\boldsymbol{c_i}$ with respect to the law $\mathcal{N}(\boldsymbol{0}, \mathrm{Diag}(\tilde{\boldsymbol{x}}))$, where $\tilde{\boldsymbol{x}}$ is a prior estimate of the flow volumes.

10.4.2 A Heuristic argument for the use of SCOD

In many cases, the $\boldsymbol{c}-$optimal design $\boldsymbol{w_c}$ is unique for every vector $\boldsymbol{c} \in \mathbb{R}^m$. In the case of the standard constraints $\sum_i w_i \leq 1$, the extended version of Elfving's theorem

for multiresponse experiments (cf. Figure 5.1, page 92) indicates that non-unicity of a $\boldsymbol{c}-$optimal design can only occur if there is a kind of parallelism between the ellipsoids $\mathcal{E}_i$ which are generated by the matrices $A_i^T A_i$ –More precisely, when a face of the generalized Elfving set is of dimension d and contains points from at least $d+2$ experiments. Although we have not proved this yet, we do not expect this to occur in the present application to telecommunications, especially when the observation matrices A_i are obtained after a left diagonal scaling which depend from previous observations of the traffic (cf. Section 10.2.2).

When the technical condition described above is fulfilled, $\boldsymbol{w_c}$ is well defined for all $\boldsymbol{c} \in \mathbb{R}^m$, and the (uniform) SCOD method is in fact a Monte Carlo approximation of the vector

$$\boldsymbol{w}^* = \mathbb{E}_{\boldsymbol{c}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_m)}[\boldsymbol{w_c}] = \int_{\boldsymbol{c}\in\mathbb{R}^m} \frac{1}{(2\pi)^{m/2}}\ \boldsymbol{w_c}\ \exp(-\frac{1}{2}\boldsymbol{c}^T\boldsymbol{c}).$$

In the remaining of this section, we say that $\boldsymbol{w}^*$ is the *ESCOD* for the experimental design problem (for Expected value of the Successive $\boldsymbol{c}-$Optimal Designs).

We now sketch a heuristic argument which establishes a relation between the ESCOD and the $A-$optimal design. The $A-$optimal design $\boldsymbol{w_A}$ minimizes $\operatorname{trace} M(\boldsymbol{w})^{-1}$ over the set $\boldsymbol{w} \in \mathcal{W} := \{\boldsymbol{w} \geq \boldsymbol{0} :\ R\boldsymbol{w} \leq \boldsymbol{d}\}$. For a random vector $\boldsymbol{c}$ following a normal distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$, we can write:

$$\begin{aligned} w_A = &\quad \operatorname*{argmin}_{\boldsymbol{w}\in\mathcal{W}}\ \operatorname{trace} M(\boldsymbol{w})^{-1}\mathbb{E}_{\boldsymbol{c}}[\boldsymbol{c}\boldsymbol{c}^T] \\ = &\quad \operatorname*{argmin}_{\boldsymbol{w}\in\mathcal{W}}\ \mathbb{E}_{\boldsymbol{c}}[\boldsymbol{c}^T M(\boldsymbol{w})^{-}\boldsymbol{c}]. \end{aligned}$$

Remarkably, if we exchange the order of the expectation and the minimization in the latter expression, we obtain the definition of the ESCOD vector $\boldsymbol{w}^*$. This, of course, does not account for a proof that the presented stochastic SCOD converges to an $A-$optimal design, since $\mathbb{E}[\cdot]$ and $\operatorname{argmin}(\cdot)$ do not commute in general. However, we observed numerically on a large number of examples that the design obtained by averaging several $\boldsymbol{c}-$optimal designs (for vectors $\boldsymbol{c_i}$ sampled from a normal distribution) was very close to the $A-$optimal design indeed. This nice property of the SCOD will be illustrated in Sections 10.4.3 and 10.5.2.

We next compare the $A-$optimal design and the ESCOD in a very simple case.

10.4.3 Comparison of the ESCOD and the A-optimal design in a simple case

In this section, we shall compute in closed form the ESCOD and compare it to the $A-$optimal designs for the case in which there are two regression vectors in $\mathbb{R}^2$:

$$\boldsymbol{a_1} = [r_1\cos(\alpha_1), r_1\sin(\alpha_1)]^T, \quad \boldsymbol{a_2} = [r_2\cos(\alpha_2), r_2\sin(\alpha_2)]^T.$$

If the vectors $\boldsymbol{a_1}$ and $\boldsymbol{a_2}$ are linearly independent, we know from Theorem 2.4.8 that the

weights of the $A-$optimal design are proportional to the square root of the diagonal of the matrix

$$B := \left(\begin{bmatrix} \boldsymbol{a_1}^T \\ \boldsymbol{a_2}^T \end{bmatrix} [\boldsymbol{a_1}, \boldsymbol{a_2}] \right)^{-1} = \begin{pmatrix} r_1^2 & * \\ * & r_2^2 \end{pmatrix}^{-1} = \frac{1}{*} \begin{pmatrix} r_2^2 & * \\ * & r_1^2 \end{pmatrix}.$$

The $A-$optimal design is thus independent from the angles α_1 and α_2:

$$\boldsymbol{w}_A = \begin{bmatrix} \frac{r_2}{r_1+r_2} \\ \frac{r_1}{r_1+r_2} \end{bmatrix}.$$

We now turn to the computation of the ESCOD $\boldsymbol{w}^*$. This design is clearly invariant to a rotation or to a scaling of the Elfving set, such that we can assume without loss of generality that $\alpha_1 = 0$ and $r_1 = 1$. In the following, we simply write r for r_2 and α for α_2 to simplify the notation. Let $\boldsymbol{c} = [\rho\cos\theta, \rho\sin\theta]^T$. The weights of the $\boldsymbol{c}-$optimal design are given by Theorem 2.4.10:

$$\begin{aligned} \boldsymbol{w}_{\boldsymbol{c}} &\propto \left| \left(\begin{bmatrix} \boldsymbol{a_1}^T \\ \boldsymbol{a_2}^T \end{bmatrix} [\boldsymbol{a_1}, \boldsymbol{a_2}] \right)^{-1} \begin{bmatrix} \boldsymbol{a_1}^T \\ \boldsymbol{a_2}^T \end{bmatrix} \boldsymbol{c} \right| \\ &\propto \begin{bmatrix} r|\sin(\alpha-\theta)| \\ |\sin\theta| \end{bmatrix}. \end{aligned}$$

After normalization, we find

$$\boldsymbol{w}_{\boldsymbol{c}} = \begin{bmatrix} \frac{r|\sin(\alpha-\theta)|}{r|\sin(\alpha-\theta)|+|\sin\theta|} \\ \frac{|\sin\theta|}{r|\sin(\alpha-\theta)|+|\sin\theta|} \end{bmatrix}.$$

Note that the $\boldsymbol{c}$-optimal design is unique for every vector $\boldsymbol{c} \in \mathbb{R}^m$, so that we can take their mean (with respect to a Gaussian distribution) without ambiguity. Since the expression of $\boldsymbol{w}_{\boldsymbol{c}}$ does not depend on ρ, the expected value $\mathbb{E}_{\boldsymbol{c}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_2)}[\boldsymbol{w}_{\boldsymbol{c}}]$ reduces to the mean of $\boldsymbol{w}_{\boldsymbol{c}}$ on the circle of radius $\rho = 1$:

$$\boldsymbol{w}^* = \frac{1}{2\pi} \int_{\theta=0}^{2\pi} \begin{bmatrix} \frac{r|\sin(\alpha-\theta)|}{r|\sin(\alpha-\theta)|+|\sin\theta|} \\ \frac{|\sin\theta|}{r|\sin(\alpha-\theta)|+|\sin\theta|} \end{bmatrix} d\theta.$$

After some (tedious !) work, we obtain $\boldsymbol{w}^* = \begin{bmatrix} u \\ 1-u \end{bmatrix}$, where

$$u = \frac{r\Big(\cos\alpha(\pi-2\alpha)(1-r^2) + (\pi r - 2\sin\alpha\ln r)(1+r^2) - 2\pi r\cos^2\alpha\Big)}{\pi\Big((1+r^2)^2 - 4r^2\cos^2\alpha\Big)}. \tag{10.12}$$

We have plotted on Figure 10.2 the difference between the first coordinates of $\boldsymbol{w}^*$ and $\boldsymbol{w}_A$, $u - \frac{r}{1+r}$, where u is the expression defined in Equation (10.12), as well as the ratio between $\phi_A(\boldsymbol{w}_A)$ and $\phi_A(\boldsymbol{w}^*)$, for different values of r and α. Note that we study the

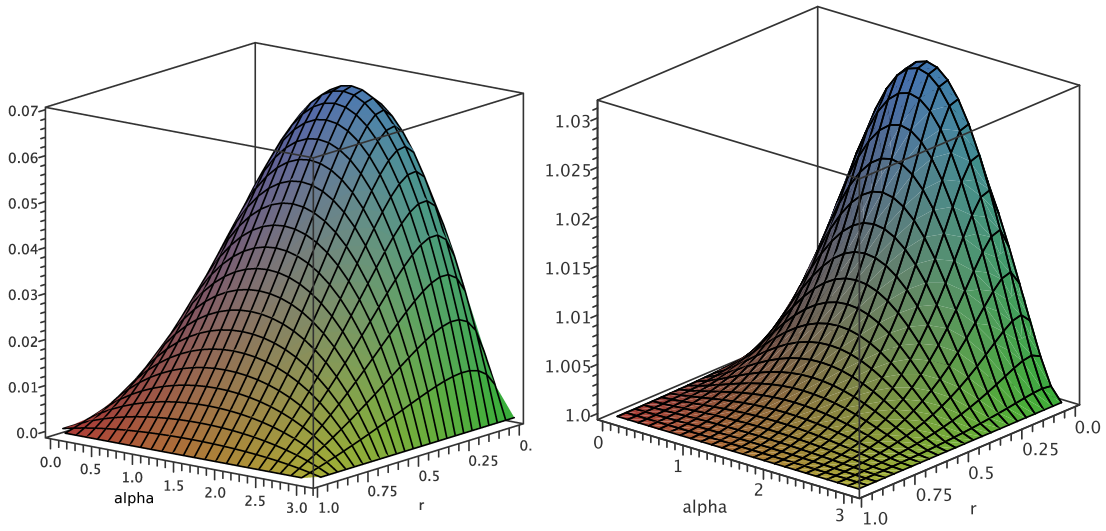


Figure 10.2: **O**ptimal design problem for the regression vectors $\boldsymbol{a_1} = [1,0]^T$ and $\boldsymbol{a_2} = [r\cos(\alpha), r\sin(\alpha)]^T$. Left: Difference between the first coordinates of the ESCOD and the $A-$optimal design: $\frac{1}{2}\|\boldsymbol{w}^* - \boldsymbol{w}_A\|_1$ **Right:** $A-$Efficiency of the ESCOD: $\phi_A(\boldsymbol{w}^*)/\phi_A(\boldsymbol{w}_A)$

effect of r in the interval $[0,1]$ only, because we can assume without loss of generality that $\|\boldsymbol{a_1}\| \geq \|\boldsymbol{a_2}\|$. On the first graph, we see that $\boldsymbol{w}^*$ is always *close* to the $A-$optimum $\boldsymbol{w}_A$. The worst case occurs for $\alpha = \pi/2, r \approx 0.17$, where the difference is of 0.069. The results are even better in terms of the $A-$efficiency achieved by the ESCOD $\boldsymbol{w}^*$: we see here that the ESCOD is always a $1.031-$approximation of the $A-$optimal design, the worst case being attained for $\alpha = \pi/2$ and $r \approx 0.09$. Remarkably, we also see on these graphs that $\boldsymbol{w}^* = \boldsymbol{w}_A$ when $r = 1$, i.e. when the two regression vectors $\boldsymbol{a_1}$ and $\boldsymbol{a_2}$ have the same length.

We noticed on this 2D-example that the situation in which the ESCOD is the furthest from the $A-$optimum is when the regression vectors are orthogonal. So we shall now study the case in which there are s orthogonal regression vectors in $\mathbb{R}^s$:

$$\forall i \in [s],\ \boldsymbol{a_i} = r_i \boldsymbol{e_i},$$

where $\boldsymbol{e_i}$ is the i^{th} vector from the canonical basis of $\mathbb{R}^s$. We denote by $\boldsymbol{r}$ the vector of the lengths of the regression vectors: $\boldsymbol{r} = [r_1, \ldots, r_s]^T \geq \boldsymbol{0}$. Similarly as in the 2D-case, we have:

$$\boldsymbol{w_A} \propto \mathbf{diag}^{1/2}\Big(\operatorname{Diag}(\boldsymbol{r})\operatorname{Diag}(\boldsymbol{r})^T\Big)^{-1} = \boldsymbol{r}^{-1},$$

where $\boldsymbol{r}^{-1}$ is the elementwise inverse of $\boldsymbol{r}$. After normalization, we find

$$\boldsymbol{w}_A = \frac{\boldsymbol{r}^{-1}}{\|\boldsymbol{r}^{-1}\|_1}.$$

The $\boldsymbol{c}-$optimal design is obtained in the same way:

$$\boldsymbol{w}_{\boldsymbol{c}} = \frac{|\operatorname{Diag}(\boldsymbol{r})^{-1}\boldsymbol{c}|}{\|\operatorname{Diag}(\boldsymbol{r})^{-1}\boldsymbol{c}\|_1},$$

and the ESCOD $\boldsymbol{w}^*$ is given by an integration on the unit sphere $\mathcal{S}^{s-1} = \{\boldsymbol{u} \in \mathbb{R}^s : \|\boldsymbol{u}\|_2 = 1\}$

$$\boldsymbol{w}^* = \int_{\boldsymbol{c}\in\mathcal{S}^{s-1}} \frac{|\operatorname{Diag}(\boldsymbol{r})^{-1}\boldsymbol{c}|}{\|\operatorname{Diag}(\boldsymbol{r})^{-1}\boldsymbol{c}\|_1} d\mu, \tag{10.13}$$

where μ is the Lebesgue measure on $\mathcal{S}^{s-1}$. We have computed some approximations of $\boldsymbol{w}^*$ for several random radius vectors $\boldsymbol{r}$ of various dimensions $s \in [100]$ thanks to Monte-Carlo simulations, in which we averaged the integrand in (10.13) for 10^5 vectors $\boldsymbol{c_i}$ drawn from a uniform distribution on $\mathcal{S}^{s-1}$. In our experiments, the efficiency ratio $\frac{\phi_A(\boldsymbol{w}^*)}{\phi_A(\boldsymbol{w_A})}$ was always smaller than 1.048, and the L_1 error $\|\boldsymbol{w}^* - \boldsymbol{w_A}\|_1$ was always smaller than 20%. This suggests that the ESCOD is a good candidate for the $A-$optimal design problem, for the case of s indepenent regression vectors in $\mathbb{R}^s$.

We point out that the previously studied experimental design problems (s independent regression vectors in $\mathbb{R}^s$) are somehow *special*, in the sense that one must select every experiment in order to obtain a full rank information matrix ($M(\boldsymbol{w}) \succ 0 \Leftrightarrow \boldsymbol{w} > \boldsymbol{0}$). We think that this situation, which we may call *low instrumented*, is close to what happens in the present industrial application, where Netflow must be activated at a significant number of locations so that $M(\boldsymbol{w})$ becomes invertible. In the *over-instrumented* situation however, the ESCOD may be irrelevant. In the quadratic regression model with 101 support points on the range $[-1, 1]$ for example, the ESCOD is quite different from the $A-$optimal design, and has a $A-$efficiency of approximately 1.175. We will investigate this feature from a theoretical point of view in future research.

10.5 Experimental results

In this section we evaluate the performance of our SCOD approach. To this end, we investigate several issues: in a first part, we compare our SCOD to the (exact) $A-$optimal design for a special instance of the problem. Then, we examine the quality of the estimation of the traffic matrix in different situations, in order to compare the performance of our method with previously proposed ones. We study separately the discrete problem of finding a subset of interfaces for Netflow, and the optimal sampling problem.

10.5.1 Data used

The data we used for those experiments comes from two networks. On the one hand, from the Abilene Internet2 backbone, which is a major academic network in the USA, and consists in $n = 11$ nodes $m = n^2 = 121$ OD pairs and $l = 50$ links (14$\times$ 2 bidirectional links,

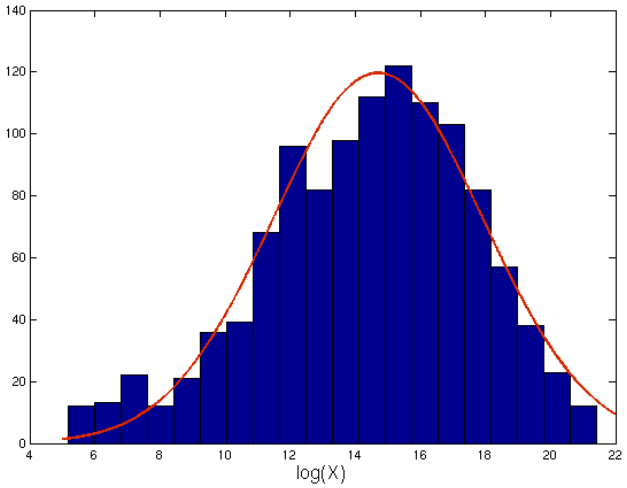


Figure 10.3: lognormal distribution of the measured flows (commercial network)

11 ingress, and 11 egress links). The topology of this network is depicted in Example 10.1.1. Real traffic matrices from this network are available through the Internet2 Observatory project. We used the measurements of the second week of April 2004, as collected by Zhang [Abi]. The data has a resolution of 10 minutes, resulting in 1008 time steps over the week.

On the other hand, we use measurements from a much larger commercial network, the international "Opentransit" network of France Telecom, which consists in $n = 116$ nodes, $m = n^2 = 13456$ OD-pairs, and $l = 436$ links. Since this network is only partially instrumented with Netflow (we dispose of Netflow measurements on 34 out of 116 routers), we simulated the missing data for the sake of experiments, by following the instructions of [NST05]. Namely, we noticed that the fit of the partially available data with a lognormal distribution was very good (see Figure 10.3), so we simulated the missing flows with respect to this distribution. Then, we assigned them to the non-measured OD pairs of the network thanks to a heuristic procedure based on the topology of the network [NST05]. The data has a resolution of 2 hours and was collected during 40 hours, so we track the flow volumes over 20 time steps.

The SNMP and Netflow measurements were simulated from the traffic matrices. The SNMP data was supposed to be almost perfect ($\sigma = 1$), and the Netflow sampling was simulated with a binomial distribution, as seen in Section 10.2.2.

10.5.2 SCOD Vs $A-$optimal designs on Abilene

We study an experimental design problem on Abilene, where the objective is to find the optimal amount of experimental effort to spend on each router (we handle the data collected on all incoming interfaces of a given router as a single experiment). Note that this setting

Design ($\times 10^{-1}$)	SCOD ($N = 10$)	SCOD ($N = 50$)	A-optimal
CPU (sec.)	3.72	18.7	492.6
w_1 (Atlanta)	0.559	0.779	0.749
w_2 (Chicago)	0.883	0.854	0.898
w_3 (Denver)	1.721	1.592	1.510
w_4 (Houston)	0.692	0.772	0.720
w_5 (Indiana)	1.458	1.291	1.361
w_6 (Kansas)	1.252	1.262	1.171
w_7 (Los Angeles)	0.556	0.572	0.657
w_8 (New York)	1.329	1.134	1.121
w_9 (Sunnyvale)	1.076	1.184	1.201
w_{10} (Seattle)	0.000	0.002	0.000
w_{11} (Washington)	0.433	0.557	0.613

Table 10.1: Abilene: comparison of the A-optimal design and SCOD.

is consistent with versions of Netflow that are earlier than the *v9* [CISb], in which setting different sampling rates on each interface of a router was not possible. In Table 10.1, we compare the $A-$optimal sampling rates found by solving the SDP (10.9), and the design obtained by the successive $\boldsymbol{c}-$optimal design approach described in Section 10.4.1. The constraint considered here was the unit cost case: $\sum_i w_i \leq 1$. The $\boldsymbol{c}-$optimal designs are computed by Program (10.10). The designs indicated in the tables were obtained by averaging $N = 10$ and $N = 50$ $\boldsymbol{c}-$optimal designs. To see the convergence of the SCOD, we have plotted in Figure 10.4 the evolution of each coordinate of the design with N.

It is striking that the designs found by these two approaches are very close and that the computation is much shorter for the SCOD. Namely, solving one instance of the $\boldsymbol{c}-$optimal problem requires only 345ms on average for this network, which is 3 orders of magnitude faster than the 514s required to solve the SDP (10.9). Furthermore, the SDP approach is intractable on large networks with more than 17 nodes.

10.5.3 Estimation methodology and Error metrics

Before studying the quality of the estimation of the traffic matrix, we describe the methodology used for the inference. For the optimal deployment problem (Section 10.2.1), we use the entropic projection approach [LTY06] (cf. Chapter 8 and 9) to track the flow volumes over time. Namely, we choose at each time step the vector of flows which is the closest to a prior (in terms of Kullback-Leibler divergence), among all the flows satisfying the measurement equation (10.2). At time $t = 1$, the prior is taken equal to the tomogravity estimate [ZRDG03] of the flows. Then, we choose as prior the previous estimate $\hat{\boldsymbol{x}}_{t-1}$. The entropic projection is carried out by the Iterated Proportional Fitting (IPF) algorithm (Algorithm 9.5.2, cf. also [LTY06]).

For the optimal sampling problem (Section 10.2.2), the observation matrix $\tilde{A}$ usually has full column rank (because Netflow is activated everywhere), such that we can use the

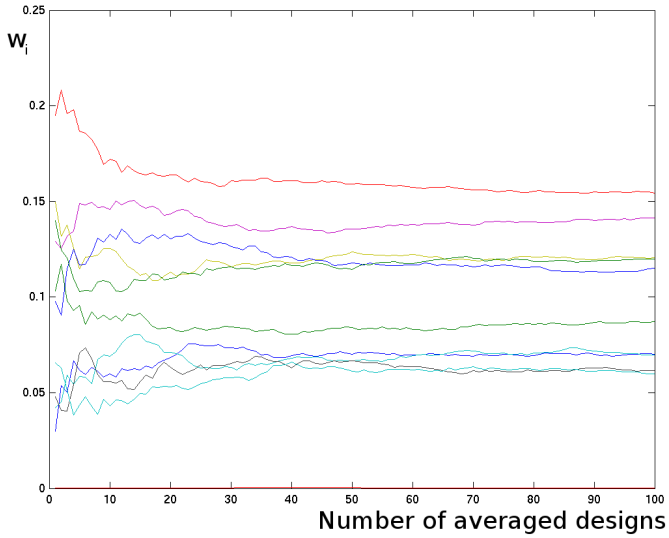


Figure 10.4: Convergence of the SCOD method. Each curve represents the evolution of a coordinate of $\boldsymbol{w}$ with the number of averaged designs.

inversion formula (10.7) to compute the best linear unbiased estimate $\hat{\boldsymbol{x}}$ of the flows, where $\Sigma(\boldsymbol{w})$ is estimated thanks to a prior estimate of the flows. To avoid eventual negative values, we next apply the IPF procedure, as in [CDVY00, ZRDG03].

To measure the quality of an estimator of the flows $\hat{\boldsymbol{x}}_t$ at a time step t, we use the classic *relative* $L2-$*error*, defined as:

$$\mathrm{Rel}_2(\hat{\boldsymbol{x}}_t) = \frac{\|\hat{\boldsymbol{x}}_t - \boldsymbol{x}_t\|_2}{\|\boldsymbol{x}_t\|_2}. \tag{10.14}$$

Similarly, the spatial distribution of the errors can be measured by the spatial relative L_2-error, which is defined for each OD flow time series $\boldsymbol{x}_{\mathrm{OD}}$:

$$\mathrm{Rel}_2(\hat{\boldsymbol{x}}_{\mathrm{OD}}) = \frac{\|\hat{\boldsymbol{x}}_{\mathrm{OD}} - \boldsymbol{x}_{\mathrm{OD}}\|_2}{\|\boldsymbol{x}_{\mathrm{OD}}\|_2}. \tag{10.15}$$

10.5.4 Netflow Optimal Deployment

We now study the case of the discrete problem presented in Section 10.2.1, where the objective is to activate Netflow only on a subset of interfaces of the network. We assume throughout this section that when Netflow is activated on an interface, it samples packets at a rate of 10^{-3}. This problem may look very academic, since routing changes occur quite often in practice, and the deployment of Netflow should not be decided in a special routing configuration. However, we show in this section that our SCOD improves on the greedy design, and we want to develop for future work a more robust version of our model. For example, if we are given several potential routing matrices – and the corresponding

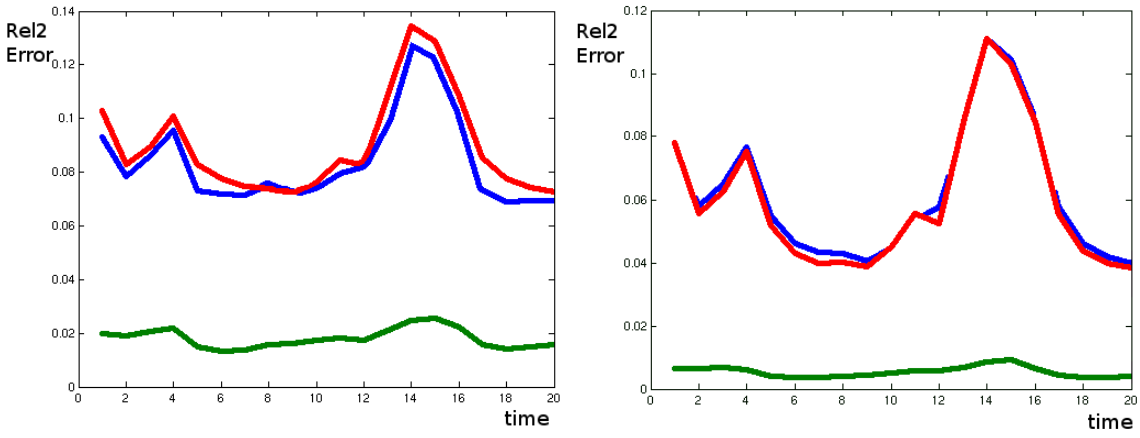


Figure 10.5: Relative L_2-error on Opentransit, for Netflow activated on 16 routers (left) and 30 routers (right), as selected by the SCOD (blue), weighted SCOD (green), and greedy "Netquest" [SQZ06] (red).

information matrices $M^{(i)}(\boldsymbol{w})-$ we could with little change write a version of our SOCP which minimizes the worst variance $\max_i \ \boldsymbol{c}^T M^{(i)}(\boldsymbol{w})^\dagger \boldsymbol{c}$. A more sophisticated idea would require the use of the *model robust* $S-$optimality criterion (cf. Sections 2.3.3 and 5.3.1).

We have plotted in Figure 10.5 the relative L_2-error of the estimate $\hat{\boldsymbol{x}}_t$ for the flow volumes over time, in two situations on Opentransit: Netflow was activated on a subset of 16 or 30 nodes, either selected by the greedy algorithm or by the SCOD procedure. We have also computed a design with a weighted SCOD procedure, in which the vectors defining the linear combinations follow $\boldsymbol{c} \sim \mathcal{N}(\boldsymbol{0}, \mathrm{Diag}(\hat{\boldsymbol{x}}))$, where $\hat{\boldsymbol{x}}$ is the tomogravity estimate of the flows at time $t = 1$ (cf. Section 10.4.1 for more details on this weighting). The number of averaged optimal designs was set to $N = 20$, so as to keep the time of computation reasonable, and because we felt that the process had almost converged.

Amazingly enough, the error of estimation is lower when the vectors $\boldsymbol{c_i}$ drawn by the SCOD procedure give more weight to large flows. While the uniform SCOD and the greedy design give results of a similar quality, the weighted SCOD substantially improves the relative L_2-error.

In order to illustrate the spatial distribution of the errors, we have plotted on Figure 10.6 the weighted quantile function of the spatial L_2-relative error: the graph indicates the fraction of traffic (on the x-axis) which is estimated with a relative L_2-error below the value on the y-axis. We see that the weighted SCOD outperforms the uniform SCOD and the greedy design for the estimation. For example, about 87% of the traffic is estimated with a relative L_2-error below 5% (for the Netflow deployment found by weighted SCOD), while this proportion falls to respectively 62% and 53% with the uniform SCOD and the greedy design. In fact, some small flows, which account for less than 1% of the total traffic, are best estimated with a uniform scheme. We have also plotted in Figure 10.7 the evolution of the L_2-relative error in function of the number of routers where Netflow is activated,

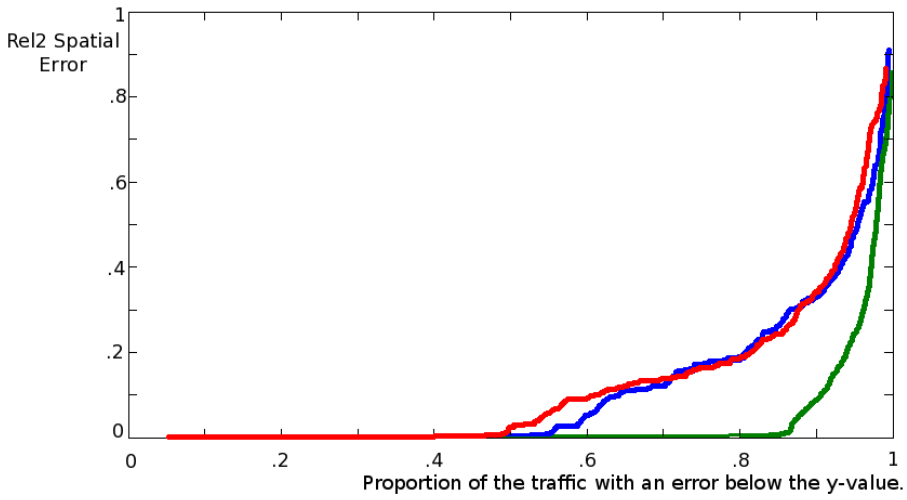


Figure 10.6: Quantile function of the spatial errors on Opentransit, for Netflow activated on 16 nodes. These nodes are selected by [SCOD (blue), weighted SCOD (green), and greedy (red)].

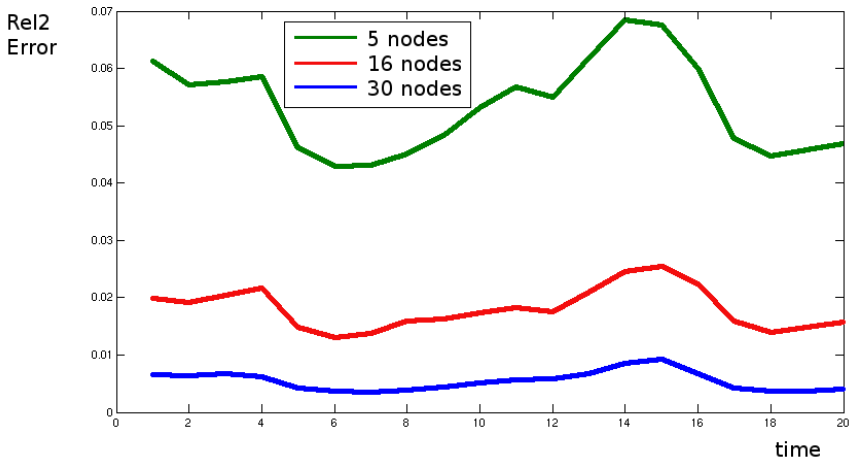


Figure 10.7: Temporal L_2-error on Opentransit, when Netflow is activated on 5, 16, and 30 nodes (selected by weighted SCOD).

to evidence the fact that a small number of Netflow measurements can yield an accurate estimation of the traffic matrix.

We show on Figure 10.9 the location of the routers found by SCOD (a), weighted SCOD (b), and the greedy algorithm (c). We notice here that the weighted SCOD procedure yields a design which is more concentrated at the "center" of the network, where the flows are probably more important. Interestingly, our problem looks somehow related to the problem of *centrality*, where the goal is to find a subset of nodes intersecting the largest possible number of shortest path in the graph. We would like to investigate this feature in a future work. We have also computed a SCOD with different weights on each link. Figure 10.8 indicates the location of the 15 links with the largest weights.

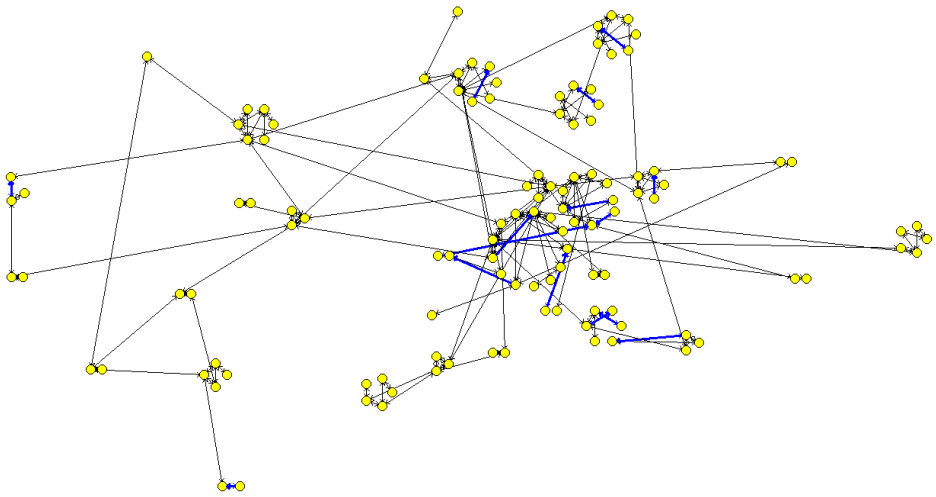


Figure 10.8: Opentransit network: Location of the 15 links with the largest weights (computed by SCOD).

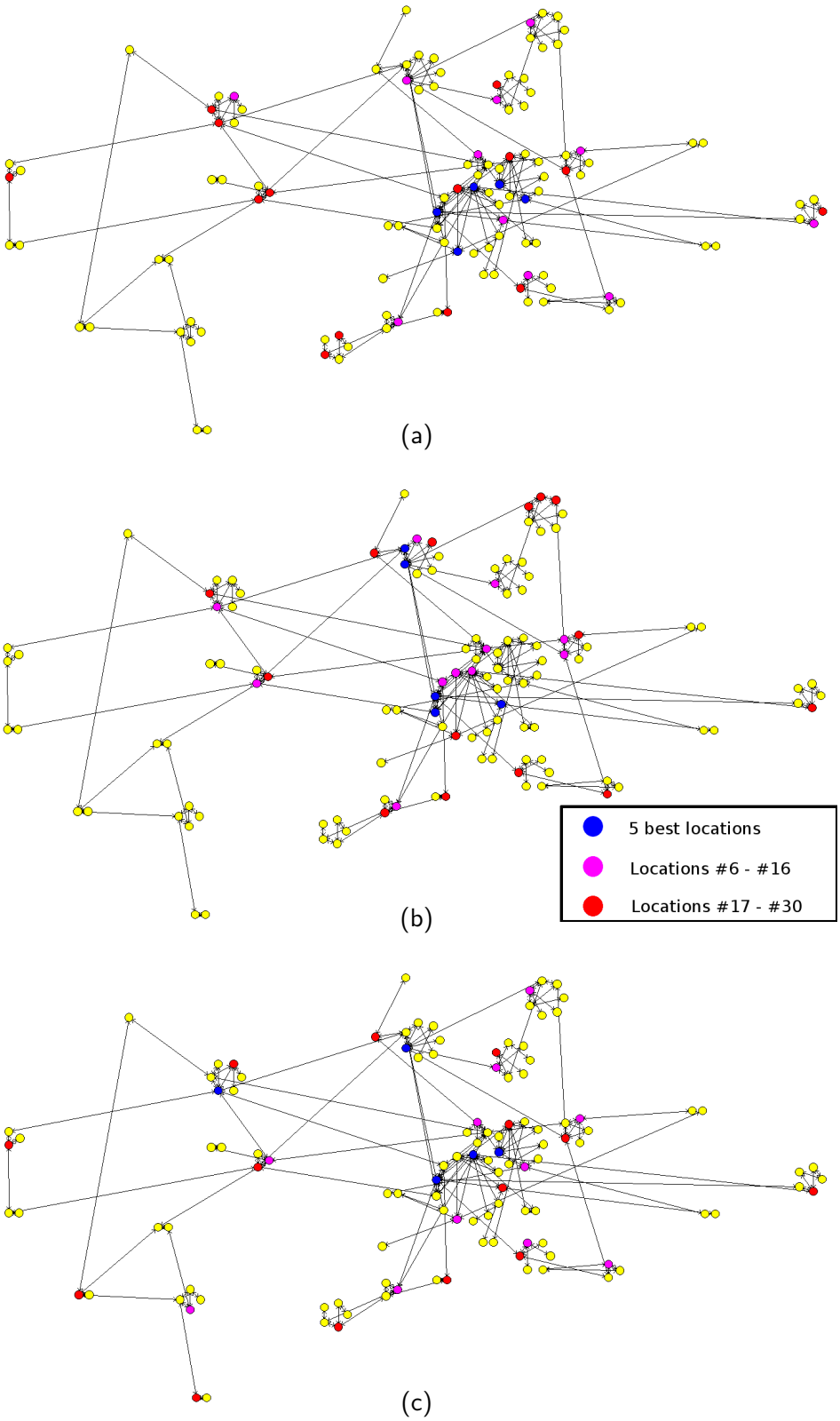


Figure 10.9: Opentransit network: Location of the routers found by SCOD (a), weighted SCOD (b), and greedy "Netquest" [SQZ06] (c). The routers in blue, purple, and red correspond to the subsets of 5, 16, and 30 routers activated in our experiments.

10.5.5 Optimal Sampling Problem

We now turn to the study of the optimization of the sampling rates for Netflow. As for the discrete problem, the SCOD are computed by averaging 20 $\boldsymbol{c}-$optimal designs, with $\boldsymbol{c} \sim \mathcal{N}(\mathbf{0}, \mathrm{Diag}(\hat{\boldsymbol{x}}))$. New sampling rates are evaluated at each time step, with the prior $\hat{\boldsymbol{x}}$ taken as the previous estimate of the flows. In a more realistic setup, we could recompute sampling rates each time a routing change occurs. In order to avoid numerical issues, we imposed a minimal sampling rate of 10^{-6} on each interface. Note that this lower bound is consistent with Cisco's Netflow manual [CISb], which specifies that the sampling rates should be set as $1/f$, where f is a parameter in the range $\{1, ..., 65535\}$.

Comparison with the Kalman filtering approach [SM08]

To illustrate that one can recover the flow volumes without any Netflow measurements on the ingress interfaces of the network (as the standard methodology suggests [FGL+01]), we have studied the case where we activate Netflow only on the 28 internal links of Abilene.

We have compared our method to the $A-$optimal design approach in a Kalman filtering context, as proposed in [SM08]. To do so, we computed sampling rates on a period of 144 time steps with this technique, using the same settings that the authors described for the case of a *noisy initialization*. For the sake of comparison, we have assumed the *unit-cost case* $\sum w_i \leq 10^{-3}$ as in [SM08], and $\boldsymbol{c}-$optimal designs are computed by Program (10.10) with $R = \mathbf{1}^T$ (the row vector of all ones) and $d = 10^{-3}$. Each SOCP was solved within roughly 0.3s with SeDuMi on a PC at 4GHz.

Figure 10.10 shows the evolution of the sampling rates on 2 interfaces of Abilene, as well as the value of the naive sampling rate ($w_{\mathrm{naive}} = 10^{-3}/s$ on every interface). Interestingly, it seems that the design computed in a Kalman filtering process *converges* to our design. This could highlight that, due to the high variability of the traffic, the prediction step $\boldsymbol{x}_{t|t-1} = C\hat{\boldsymbol{x}}_{t-1}$ (with $C = \boldsymbol{I}_m$ as in [SM08]) of the Kalman filter is of poor quality compared to the correction step which uses the Netflow measurements. Moreover, the flows computed by our approach have a relative L_2 error in the order of 10^{-3}, while the error attains 20% with the Kalman filter: Of course, the huge difference between these results does not come from the sampling rates, which are quite similar, but only from the estimation methodology: Simply inverting the sampling measurements (as we do) yields better results than processing them in a Kalman filter, since the state transition equation from one time step to the next one may be inaccurate.

We still want to evaluate the benefits of considering the past measurements for the computation of the sampling rates. So we built a new estimate of the flows, where the Kalman filter is used only to update the constant term of the covariance matrix which accounts for the past measurements. Then, the sampling rates are selected so as to *minimize* this covariance matrix. To speed up the computation –which is very expensive because the matrix accounting for the information on the past measurements is typically full– we used

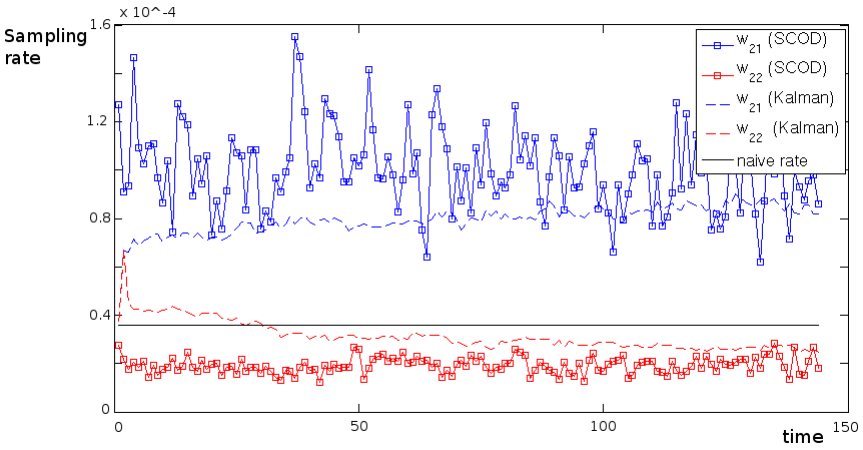


Figure 10.10: Evolution of the sampling rates on 2 interfaces of Abilene.

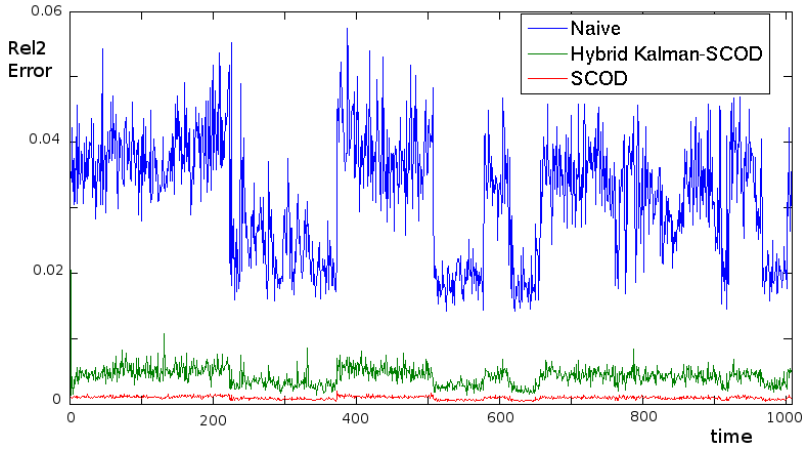


Figure 10.11: Relative $L2$ error for different sampling rates of Netflow, on Abilene.

our SCOD scheme in place of the $A-$optimality SDP (2.18). Finally, the estimation is carried out by the inversion Formula (10.7) and the IPF. We compare in Figure 10.11 the relative $L2$ error of this new estimate (called "Hybrid Kalman-SCOD") with the estimations of the flows based on the naive sampling rates and the SCOD. Our sampling rates perform much better than the naive ones. Note that the estimation of the flows is very accurate with our sampling rates, although no Netflow measurement was performed on the ingress links.

It is also clear that taking into account the past measurements does not yield any improvement on this example. In a Kalman filtering context, a given interface might not have a high rate during two successive time steps, since some information on the flow at time $t-1$ is used for the estimation at time t, through the transition equation $X_{t|t-1} = CX_t$. However, the lack of accuracy of this model makes it useless to take into account this prior information, and it is better to keep high sampling rates at all time on valuable measurements.

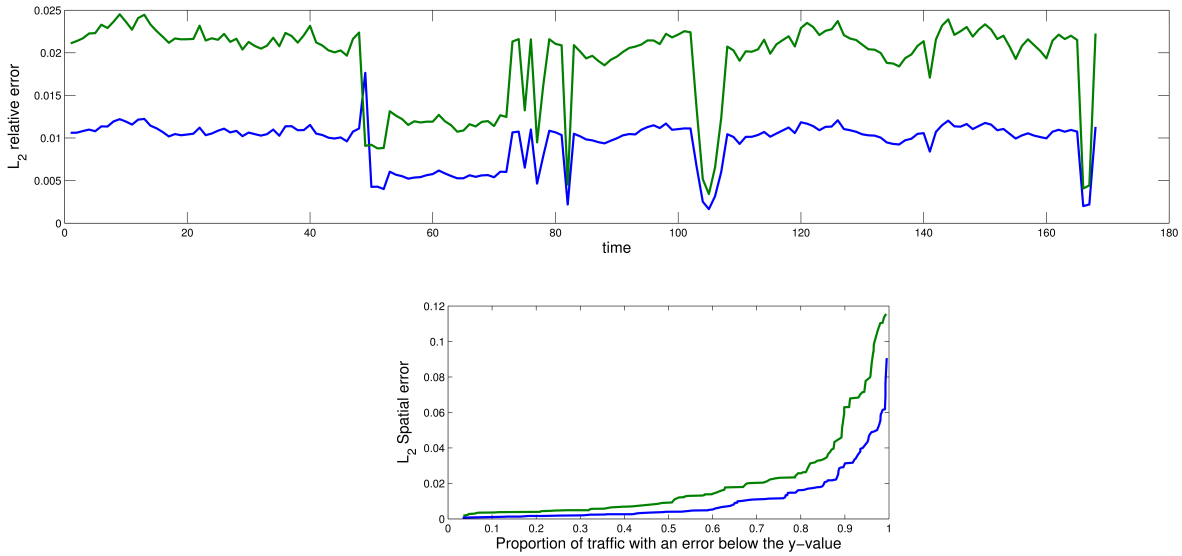


Figure 10.12: Temporal evolution and Spatial distribution of the spatial relative L_2 error for the optimal sampling problem with per-router constraints on Abilene. In blue: SCOD sampling rates. In green : Naive (equal rates on every incoming interfaces of each router).

Per-router optimization

Let us now turn to the case of the more realistic, per-router constraints described in Section 10.2.3. The problem is to select the sampling rates on each incoming interface of every router, when a target overhead is given (the maximal number of packets to be analyzed by Netflow at each router location). For this experiment, we have set a target of 10^6 packets to be analyzed by each router during an observation period of one hour. With this setting, the sampling rates returned by the SCOD procedure were typically in the range $[10^{-6}, 10^{-4}]$. We have also built the following naive sampling rates in order to evaluate the benefits of our method: the values are selected so that the sampling rates are the same on each incoming interface of a given router. Assume for example that the router $\mathcal{R}$ receives some data from its set of incoming interfaces $\mathcal{I}_\mathcal{R}$, and the volume of traffic on the link $i \in \mathcal{I}_\mathcal{R}$ is f_i. With the model described in Section 10.2.3, the row corresponding to router $\mathcal{R}$ in the constraint equation $R\boldsymbol{w} \leq \boldsymbol{d}$ is thus

$$\sum_{i \in \mathcal{I}_\mathcal{R}}^{k} f_i w_i \leq 10^6,$$

and we set the naive rates $\forall i \in \mathcal{I}_\mathcal{R},\ w_i = 10^6 / \sum_{j \in \mathcal{I}_\mathcal{R}} f_j$. We have plotted in Figure 10.12 the temporal evolution and the spatial distribution of the errors of estimation with these sampling rates. During every observation period, the estimation error is roughly two times smaller with the optimized sampling rates. Concerning the spatial distribution of the errors, we see e.g. that roughly 65% of the traffic is estimated with a relative L_2-error below 1% with the optimized sampling rates, while this number falls to 45% with the naive rates.

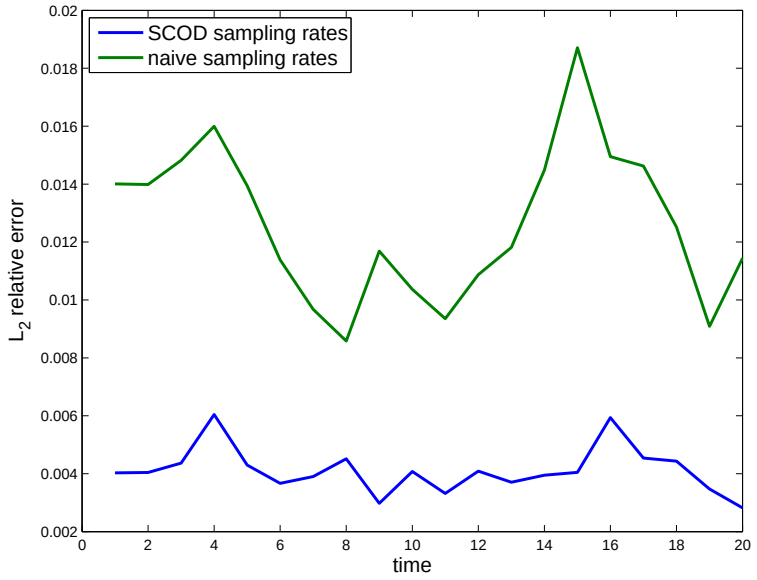


Figure 10.13: Relative L_2-error Vs time, for the per-router sampling problem on Opentransit.

We have also performed the same experiment on Opentransit. The computational effort becomes an issue at this scale, since 3 minutes were required to compute a single $\boldsymbol{c}-$optimal design (see also Table 6.3 at page 117). In consequence, if we want to take an average of 20 such optimal designs, one hour of computation is required. Note however that this task is highly parallelizable (if one disposes of N processors, then N $\boldsymbol{c}-$optimal designs can be computed simultaneously). The relative L_2-error of the flows is plotted on Figure 10.13, for the flows estimated from measurements with naive and SCOD sampling rates.

Chapter 11

Perspectives for a better spatio-temporal modelling of traffic matrices

Many methods presented in Chapter 8 for the estimation of traffic matrices rely on a temporal model for the dynamic of the TM (Kalman with a diagonal transition matrix C [CVFC09], PAMTRAM [LTY06]), or a spatial model of the TM (EM algorithm [CDVY00], fanouts [PTL04], splines [CVFC09]), or both (Kalman with an arbitrary transition matrix C [SSNT05], PCA [SLT+05]). Apart from the fanouts method, the spatial modelling carried out in the latter approaches concerns the vectorized traffic matrix $\boldsymbol{x}$ and assumes the mutual independence of the OD flows.

However, this simple independence model is not fully satisfactory on a practical point of view. For example, assume that some long awaited content becomes available on the Internet, *close* to an access point d. It is likely that users from many places will request to access this content, thus inducing a peak of traffic from many nodes in the network to d. This example shows that two OD pairs sharing a common destination are likely to be correlated. We shall evidence this fact in Section 11.1, by studying the low rank structure of real traffic matrices.

We will next try to handle the spatial and temporal correlations simultaneously, thanks to the theory of low rank tensor decompositions (Section 11.2). We shall present this framework, and study the potential of tensor methods by decomposing real traffic matrices. Finally, we shall present some suggestions for future estimation techniques relying on tensors.

11.1 Low rank structure of traffic matrices

In this section, we study the spectrum of real traffic matrices, and we propose a spectral model. This is a preliminary work which has not given birth to a practical method for the

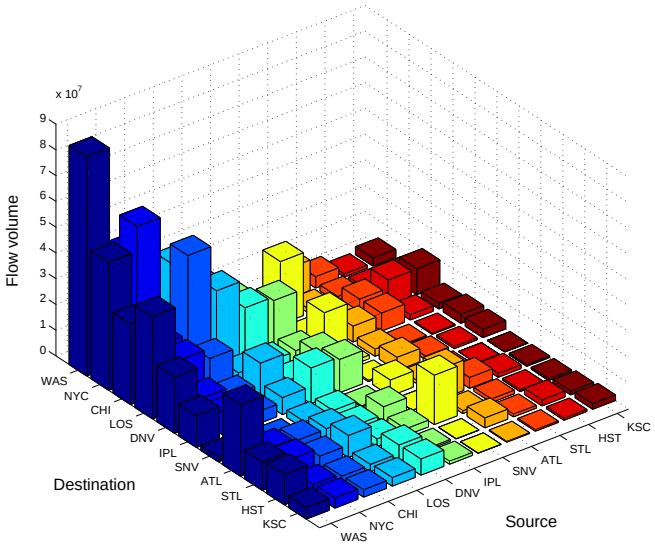


Figure 11.1: Sample snapshot of an Abilene TM

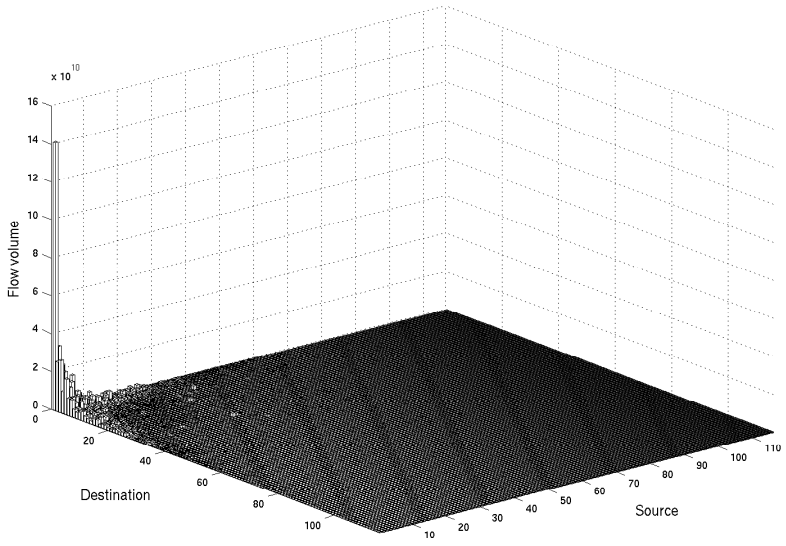


Figure 11.2: Sample snapshot of an Abilene TM

estimation of traffic matrices yet, and thus we present it in the *perspectives* chapter. We think that the low rank structure of traffic matrices is a very important property, and it justifies the tensor approach of the next section.

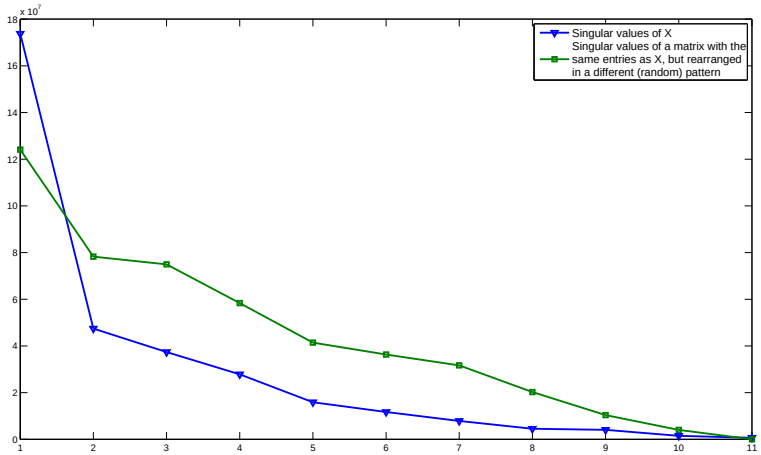


Figure 11.3: Spectrum of an Abilene TM

11.1.1 Spatial correlations

In this chapter, we do not vectorize traffic matrices anymore. Therefore, each snapshot of a TM on a network with n nodes is arranged as an $n \times n$ matrix, the $(o, d)-$entry of which indicates the volume of traffic from o to d (during the considered time interval). We have plotted on Figure 11.1 (resp. Figure 11.2) a sample snapshot X_t of a traffic matrix from Abilene (resp. Opentransit). Both plots show a strong correlation between sources and destinations. In particular, the rows of the the Abilene TM seem to be roughly mutually proportional, and the same observation holds for its columns. This indicates that the traffic matrix X_t can be well approximated by a rank-one matrix $X_t = \boldsymbol{u}\boldsymbol{v}^T$, which is consistent with the popular use of the gravity prior (cf. Chapter 9).

To verify this statement, we have plotted the spectrum of these Abilene and Opentransit TMs on Figures 11.3 and 11.4, respectively. The singular values decrease quickly in both graphs, showing that most of the *energy* from these traffic matrices can be captured by a low rank approximation (especially for Opentransit). Recall that the traffic is lognormally distributed among the OD pairs (as noticed by Nucci, Sridharan and Taft [NST05], cf. Section 10.5.1), such that the traffic matrices are approximately sparse. However, matrices with log-normally distributed entries are not necessary of low rank. In Figures 11.3 and 11.4, we have also plotted the spectrum of a matrix with the same entries as the original TM, but permuted in a random way; we see in both cases that the real traffic matrix has a much lower *dimensionality* than the permuted one. In consequence, the (approximate) low rank structure is not an artefact due to the pseudo-sparsity of the TM, and reflects the spatial correlations between sources and destinations.

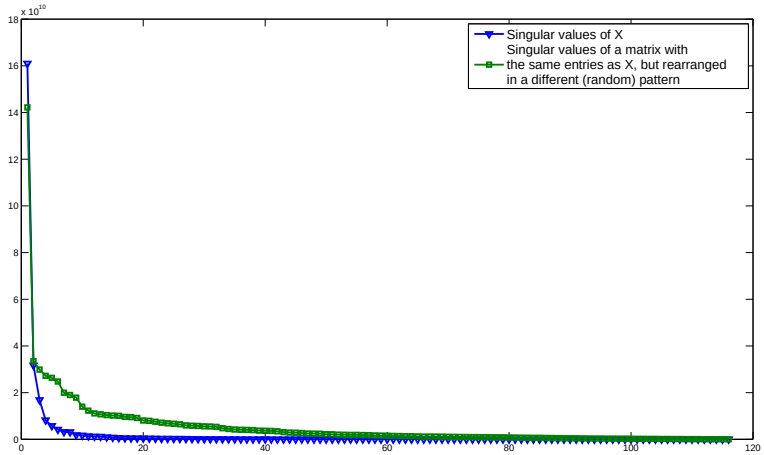
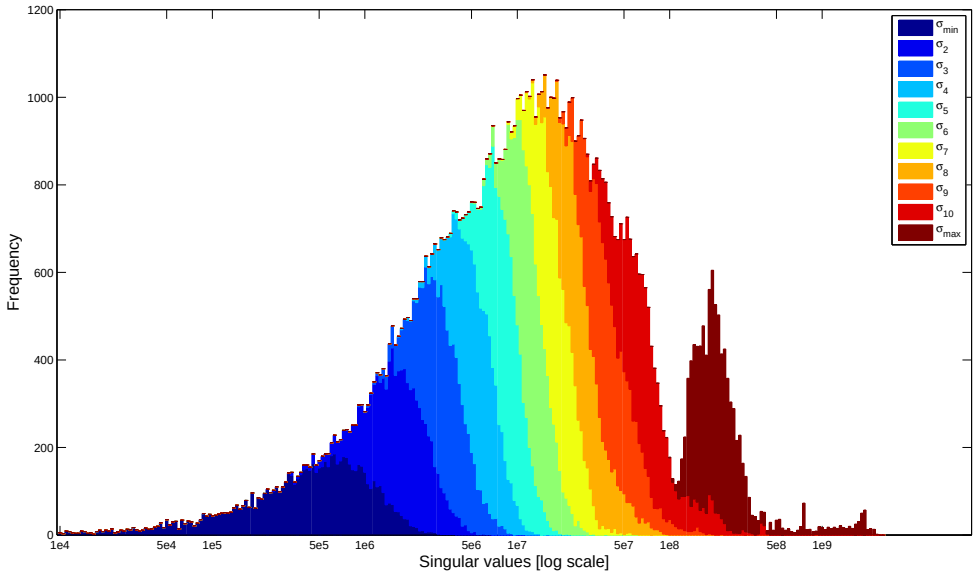


Figure 11.4: Spectrum of an Opentransit TM

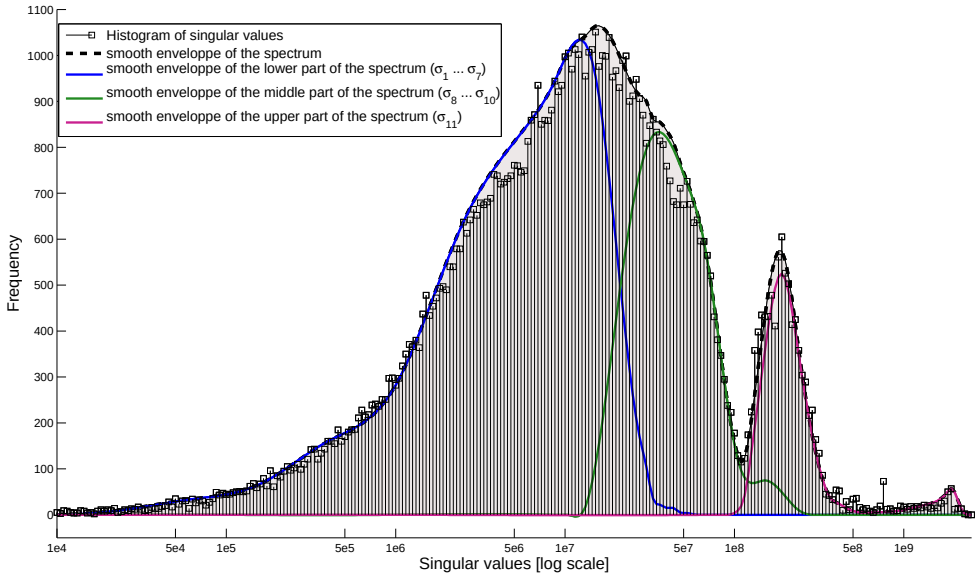
11.1.2 A statistical model for the error matrix

A possible starting point to build a spatial model for the traffic matrices is to study the empiric distribution of their spectrum. To do this, we have computed the SVD of 8064 Abilene traffic matrices, as corresponding to four complete weeks of Netflow measurements with a resolution of 5 minutes. The histogram of Figure 11.5(a) shows, on a logarithmic scale, the distribution of the singular values of these traffic matrices, by detailing the contribution of each ordered singular value to the whole distribution. On the same figure, Graph (b) separates the singular values distribution in three parts, corresponding to the lower, middle, and upper part of the spectrum. It is striking that this distribution exhibits two distinct *humps*, one accounting for the leading singular value $\sigma_{\max}$ and the other one for the remaining singular values.

This superposition of spectrums let us think that any traffic matrix X_t can be decomposed as the sum of a *deterministic, low rank matrix* X_R (with singular values in the upper part of the distribution), plus a *noise matrix* E with smaller singular values. Finding the number of eigenvalues which constitute the deterministic and the noisy part of the traffic matrix is a subtle task. For the Abilene Network, four eigenvalues for the deterministic part, and seven for the noise might be a good trade-off. We have computed the best rank-R approximation (by truncating the SVD) of one week of Abilene traffic matrices X_t with a resolution of one hour (168 traffic matrices). We have sorted the $m = 121$ OD flows in three categories: 36 *large flows* (75% of the traffic), 44 *medium flows* (20% of the traffic), and 41 *small flows* (the remaining 5%). The value $R = 4$ is the smallest for which the average of the spatial relative L_2-error (defined in Equation 10.15) of the large flows is less than 10% (this average error rises to 28% for the medium flows, and to 105% for the small ones). We have plotted on Figure 11.6 the evolution of the flow volumes of three OD pairs (one small, one medium and one large), as well as the approximation by the truncated



(a)



(b)

Figure 11.5: Distribution of the singular values of 8064 Abilene TMs

SVD (with $R = 4$) and the error. While the estimation is excellent for the large flow, and quite good for the medium one, we see that the the best rank-4 approximation still misses a significant part of the information on the small flow (on the right of the figure), which has a mean volume 20 times smaller than the large flow, since the error is periodic.

Another reason for the choice of $R = 4$ comes from the theory of *random matrices*. If

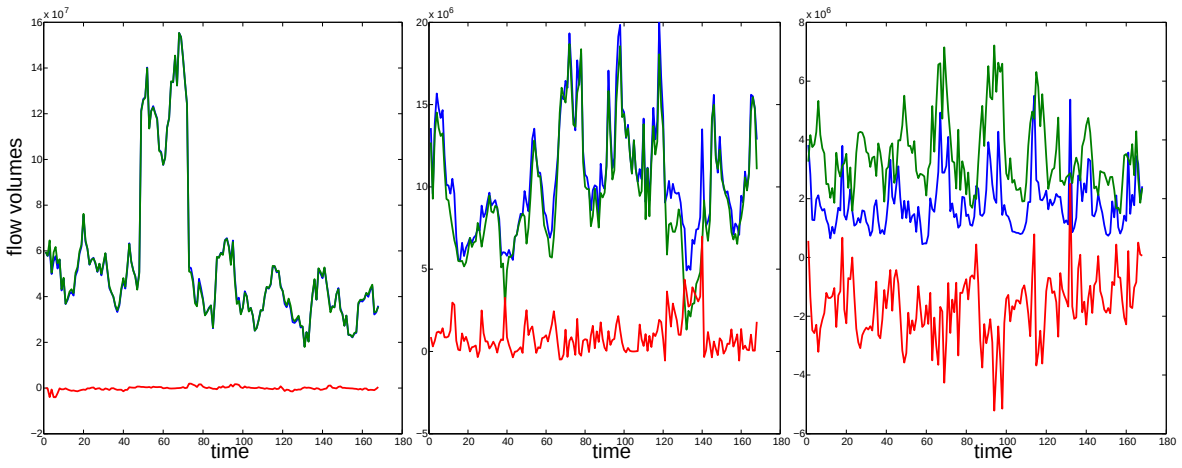


Figure 11.6: In blue: temporal evolution of a large flow (right), a medium flow (middle) and a small flow (right). In green: evolution of the flow volumes estimated by rank-4 approximations of the TMs. In red: error of estimation.

the vectors $\boldsymbol{z_1}, \ldots, \boldsymbol{z_p} \in \mathbb{R}^n$ are iid distributed with $\mathcal{N}(\mathbf{0}, \Sigma)$, we say that the $n \times n$ matrix $Z = \sum_{i=1}^{p} \boldsymbol{z_i} \boldsymbol{z_i}^T$ follows a Wishart distribution with p degrees of freedom and covariance Σ, and we note

$$Z \sim \mathcal{W}_n(p, \Sigma).$$

For small values of R, the shape of the distribution of the $11 - R$ smallest singular values of the traffic matrix (see e.g. the blue curve in Figure 11.5(b)) looks similar to the distribution of the eigenvalues of a Wishart matrix. So we tried to fit this curve to the theoretical p.d.f. of the spectrum of $\mathcal{W}_n(p, \sigma^2 \boldsymbol{I})$ for several values of R, p, and σ. The best fit was obtained for $R = 4$, $p = 7$, and $\sigma^2 \approx 6 \cdot 10^5$. In Figure 11.7, we have plotted both the histogram of the 7 smallest singular values of the TM, as well as the p.d.f. of the eigenvalues of $\mathcal{W}_{11}(7, \sigma^2 \boldsymbol{I})$, and three sample spectrums of traffic matrices. The rays out of the Wishart theoretical distribution represent the part of the traffic matrix which carries information.

A plausible model for E is one in which the law of the singular values is the same as if they were generated from a Wishart distribution. If we expect the error matrix to be invariant under orthogonal transformations, we can think of modelling the rank $n - R$ error matrix E as follows:

$$E = U \operatorname{Diag}(\boldsymbol{\lambda}) V^T,$$

where $\boldsymbol{\lambda}, U$ and V are independent, $\boldsymbol{\lambda}$ is a random vector which follow the joint distribution of the unordered eigenvalues of $\mathcal{W}_n(n - R, \sigma^2 \boldsymbol{I})$, and U and V are $n \times n$ Haar distributed matrices (see the definition below). The distribution of the $n - R$ nonzero entries of $\boldsymbol{\lambda}$ is the same as the joint distribution of the eigenvalues of a matrix from $\mathcal{W}_{n-R}(n, \sigma^2 \boldsymbol{I})$, which

is known (see e.g. [Jam64, Ede89, ER05]). The joint p.d.f of $\boldsymbol{\lambda}$ is:

$$f(\boldsymbol{\lambda}) = \kappa \prod_{i=1}^{n-R} \left(\frac{\lambda_i}{\sigma^2}\right)^{\frac{R-1}{2}} \exp\Big(-\sum_{k=1}^{n-R} \lambda_k/(2\sigma^2)\Big) \prod_{1\le i<j\le n-R} \frac{|\lambda_i - \lambda_j|}{\sigma^2},$$

where the normalizing constant

$$\kappa = 2^{-n(n-R)/2} \prod_{j=1}^{n-R} \frac{\Gamma(3/2)}{\Gamma(1+j/2)\Gamma(R/2+j/2)}.$$

We say that a matrix U is Haar distributed if U is drawn from a uniform distribution in the *Stiefeld manifold* $\{V \in \mathbb{R}^{n\times n} : V^T V = \boldsymbol{I}\}$. Note that a simple way to generate a Haar matrix is to compute the QR factorization of a random $n \times n$ matrix with Gaussian iid entries. If Q is normalized such that the diagonal elements of R are positive, then Q is a Haar matrix [Ste80].

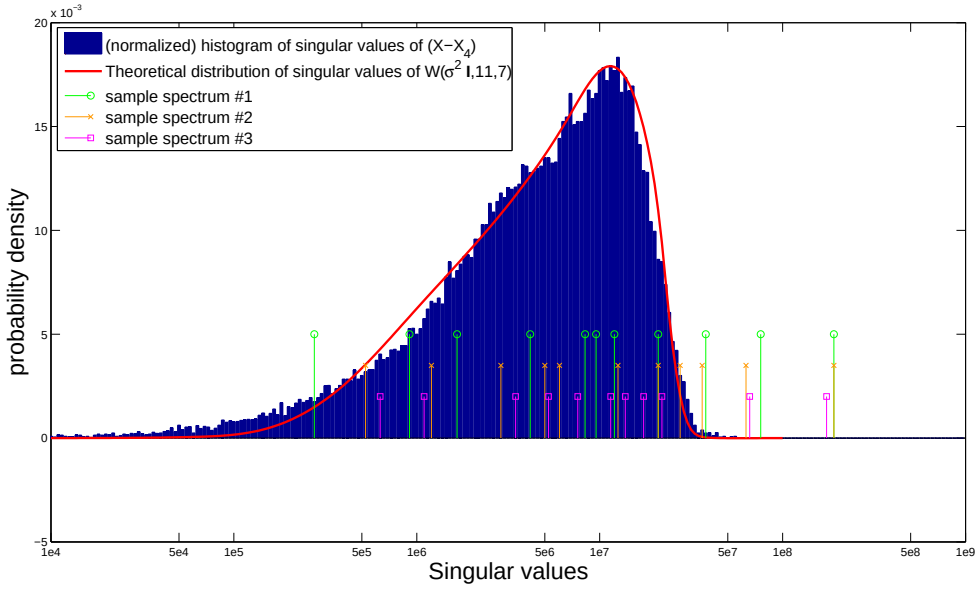


Figure 11.7: Histogram of the 7 smallest singular values of 8064 real Abilene TM, together with the theoretical distribution of the singular values of a Wishart matrix with 7 degrees of freedom. We have also plotted three sample spectrums. We think that the singular values in the *hump* of the distribution correspond to noise, while the outermost bars represent the *deterministic* part of the TM.

This proposition of model is far from providing a method for the estimation of the TM from partial measurements. However, we think that modelling the low rank structure of TMs is essential. The fit of the lower part of the spectrum of the TMs with that of a Wishart distribution indicates that it can be considered as *noise* indeed. The relevant information in the traffic matrices is mainly carried by a few eigenvectors. A related study on financial correlations was done by Laloux, Cizeau, Bouchaud and Potters [LCPB00].

The work of Zhang, Roughan, Willinger and Qiu [ZRWQ09] showed the usefulness of low rank models for the $m \times T-$ dynamic traffic matrix X (i.e. spatio-temporal correlations are considered, but no origin-destination correlations). They proposed indeed to search a low rank traffic matrix satisfying the observation equations, and noticed that this model yields a clear improvement with respect to the *raw* tomogravity estimate, when a small number of Netflow measurement are allowed. Our work suggests that it might be useful to consider origin-destination correlations as well. The mathematical objects which are best suited to handle muti-mode correlations are the *tensors*. We study the potential of tensor decompositions in the next section.

11.2 Low rank decompositions of real *traffic tensors*

When seen over time, the traffic matrix is in fact a tri-dimensional object: origine $\times$ destination $\times$ time. We shall denote by $\mathcal{X}$ this tri-dimensional array with triple indexed components $x_{o,d,t}$. In what follows, we shall abusively refer to such multi-dimensional arrays as *tensors*. In fact, the correct definition of a (covariant) tensor of order 3 is a multi-linear application:

$$\mathcal{T} : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \mathbb{R}^{d_3} \mapsto \mathbb{R}.$$

This application can be represented as a three-way array $\left(t_{i,j,k}\right)_{i\in[d_1],\ j\in[d_2],\ k\in[d_3]}$ which describes the action of $\mathcal{T}$ on the canonical basis of $\mathbb{R}^{d_1}$, $\mathbb{R}^{d_2}$ and $\mathbb{R}^{d_3}$. Note however that we would obtain a different array $(t'_{i,j,k})$ if we choose some other basis for $\mathbb{R}^{d_1}$, $\mathbb{R}^{d_2}$ and $\mathbb{R}^{d_3}$.

While low rank approximations of matrices is a completely understood problem (through the singular value decomposition), the low rank approximation of tensors is an active research topic. For more than 40 years, many theoretical and computational work has been done to build an analog of the singular value decompositon to tensors. The idea is to decompose a tensor as the sum of a few *low rank tensors*, capturing as much *energy* as possible from the original tensor.

11.2.1 Tensor decompositions

In this section, we introduce the relevant notation for tensors, and we briefly summarize some results and algorithms for tensor decompositions. For simplicity, we limit ourselves to third order tensors, but the results below can of course be generalized to higher orders. We refer the reader to the recent review of Kolda and Bader [KB09] for more details.

Some notation

In the sequel, we use calligraphic letters for third-order tensors ($\mathcal{A}, \mathcal{B}, \ldots$). The element (i, j, k) of $\mathcal{X}$ is denoted by $x_{i,j,k}$; the k^{th} column of a matrix U is denoted by $\boldsymbol{u_k}$. The

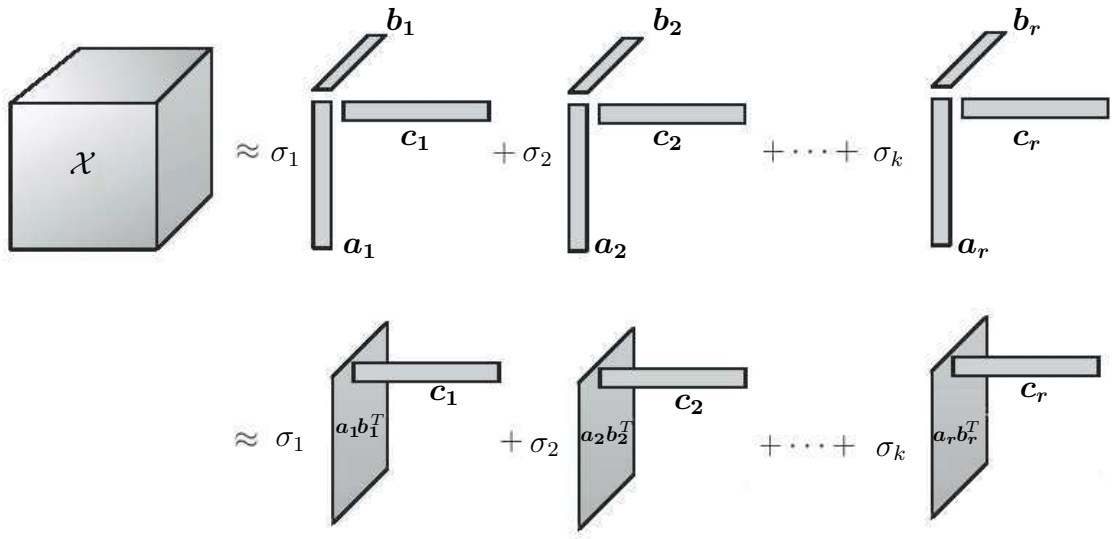


Figure 11.8: Rank-r approximation of $\mathcal{X}$. Parts from this figure are from [KB09].

Frobenius norm of a tensor $\mathcal{X} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is

$$\|\mathcal{X}\|_F := \sqrt{\sum_{i \in [d_1]} \sum_{j \in [d_2]} \sum_{k \in [d_3]} x_{i,j,k}^2}.$$

We denote by $\circ$ the vector outer product. A tensor $\mathcal{X} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is said to be *of rank one* if it can be written as $\mathcal{X} = \boldsymbol{a} \circ \boldsymbol{b} \circ \boldsymbol{c}$ for some vectors $\boldsymbol{a} \in \mathbb{R}^{d_1}$, $\boldsymbol{b} \in \mathbb{R}^{d_2}$, and $\boldsymbol{c} \in \mathbb{R}^{d_3}$, such that its element $x_{i,j,k} = a_i b_j c_k$. By analogy to matrices, the rank of a tensor $\mathcal{A}$ is defined as the smallest number r such that $\mathcal{A}$ may be decomposed as the sum of r rank-one tensors.

CP decomposition

The problem of finding the matrix of rank r which best approximates a given matrix is known to have a simple solution provided by the truncated SVD. Let $U\Sigma V^T$ be the singular value decomposition of a $d_1 \times d_2-$matrix X of rank R: U and V are matrices with R orthonormal columns of respective dimension d_1 and d_2, and Σ is diagonal with nonzero entries $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_R$. We can write $X = \sum_{k=1}^{R} \sigma_k \, \boldsymbol{u_k} \boldsymbol{v_k}^T$, or with the outer product notation: $X = \sum_{k=1}^{R} \sigma_k \, \boldsymbol{u_k} \circ \boldsymbol{v_k}$. Then, the matrix X_r of rank r which best approximates X (for the Frobenius norm) is given by

$$X_r = \sum_{k=1}^{r} \sigma_k \, \boldsymbol{u_k} \circ \boldsymbol{v_k}.$$

Several problems occur when trying to generalize the SVD to tensors. The natural idea

would be to decompose rank-R tensors under a form

$$\mathcal{X} = \sum_{k=1}^{R} \sigma_k \, \boldsymbol{u_k} \circ \boldsymbol{v_k} \circ \boldsymbol{w_k}, \tag{11.1}$$

such that the columns of $U, V,$ and W are orthonormal. However, the decomposition of a rank-R tensor as the sum of R rank-one tensors is often unique (up to a permutation of the rank-one terms), which prevents one from imposing the orthonormality of the $\boldsymbol{u_k}$, $\boldsymbol{v_k}$ and $\boldsymbol{w_k}$. A consequence is that the truncation of Decomposition (11.1) to the r dominant terms does not coincide with the best approximation of rank r. Worse than that, the r terms in the best rank-r approximation of $\mathcal{X}$ may be completely different from the r dominant terms in the best rank-$(r+1)$ approximation. In consequence, a sequential scheme in which, at stage j, one adds the best possible rank-one tensor to the previously computed best rank-$(j-1)$ tensor is suboptimal.

The problem is thus to find simultaneously the r terms of the best rank-r approximation (cf. Figure 11.8):

$$\begin{aligned} \min_{\boldsymbol{\sigma}, U, V, W} \quad & \|\mathcal{X} - \hat{\mathcal{X}}\|_F \\ \text{s.t.} \quad & \hat{\mathcal{X}} = \sum_{k=1}^{r} \sigma_k \, \boldsymbol{u_k} \circ \boldsymbol{v_k} \circ \boldsymbol{w_k}. \\ & \forall k \in [r], \quad \|\boldsymbol{u_k}\| = \|\boldsymbol{v_k}\| = \|\boldsymbol{w_k}\| = 1. \end{aligned} \tag{11.2}$$

The variables in this problem are $\boldsymbol{\sigma} \in \mathbb{R}^r$ and the matrices U, V, and W of respective size $d_1 \times r$, $d_2 \times r$ and $d_3 \times r$. We can also use the standard double brackets notation for the rank-r decomposition of $\hat{\mathcal{X}}$ in Problem (11.2):

$$\hat{\mathcal{X}} = [\![\boldsymbol{\sigma}; U, V, W]\!] := \sum_{k=1}^{r} \sigma_k \, \boldsymbol{u_k} \circ \boldsymbol{v_k} \circ \boldsymbol{w_k}.$$

This approximation is usually called *CANDECOMP/PARAFAC*, or *CP* for short, after the *canonical decomposition* of Harshman [Har70] and the *parallel factors* of Carroll and Chang [CC70].

The optimization problem (11.2) is nonconvex. Harshman [Har70] and Carroll and Chang [CC70] have proposed independently an alternating least square (ALS) procedure to approximate its solution, which can be seen as a generalization of the *power method* to tensors. The principle is to solve iteratively Problem (11.2) for U (with fixed V and W), then for V (with fixed fixed U and W), etc. The ALS procedure has been summarized in Algorithm 11.2.1, where the following notation have been used: $X_{(i)}$ is the $d_i \times \prod_{j \in [3],\ j \neq i} d_j -$matrix representing the tensor $\mathcal{X}$, *unfolded along the* i^{th} *dimension*. The $d_1 d_2 \times k -$matrix $A \circledast B$ is the Khatri-Rao product of $A = [\boldsymbol{a_1}, \ldots, \boldsymbol{a_r}] \in \mathbb{R}^{d_1 \times r}$ and $B = [\boldsymbol{b_1}, \ldots, \boldsymbol{b_r}] \in \mathbb{R}^{d_2 \times r}$,

Algorithm 11.2.1 CP: Alternating least squares (ALS)

Input: $\mathcal{X}, r, \epsilon$.
Initialize $\boldsymbol{\sigma} \in \mathbb{R}^r$, $U \in \mathbb{R}^{d_1 \times r}$, $V \in \mathbb{R}^{d_2 \times r}$, and $W \in \mathbb{R}^{d_3 \times r}$ in some way.
repeat
  Fix $(\boldsymbol{\sigma}, V, W)$, and solve Problem (11.2) in U:
    $U \leftarrow X_{(1)}(W \circledast V)(WW^T \odot VV^T)^\dagger$;
  Normalize the columns of U:
    $\boldsymbol{\mu} \leftarrow \mathbf{diag}(U^T U)$; $\boldsymbol{\sigma} \leftarrow \boldsymbol{\sigma} \odot \boldsymbol{\mu}$; $U \leftarrow \mathrm{Diag}(\boldsymbol{\mu})^{-1} U$;

  Fix $(\boldsymbol{\sigma}, U, W)$, and solve Problem (11.2) in V:
    $V \leftarrow X_{(2)}(W \circledast U)(WW^T \odot UU^T)^\dagger$;
  Normalize the columns of V:
    $\boldsymbol{\mu} \leftarrow \mathbf{diag}(V^T V)$; $\boldsymbol{\sigma} \leftarrow \boldsymbol{\sigma} \odot \boldsymbol{\mu}$; $V \leftarrow \mathrm{Diag}(\boldsymbol{\mu})^{-1} V$;

  Fix $(\boldsymbol{\sigma}, U, V)$, and solve Problem (11.2) in W:
    $W \leftarrow X_{(3)}(V \circledast U)(VV^T \odot UU^T)^\dagger$;
  Normalize the columns of W:
    $\boldsymbol{\mu} \leftarrow \mathbf{diag}(W^T W)$; $\boldsymbol{\sigma} \leftarrow \boldsymbol{\sigma} \odot \boldsymbol{\mu}$; $W \leftarrow \mathrm{Diag}(\boldsymbol{\mu})^{-1} W$;

  $\hat{\hat{\mathcal{X}}} \leftarrow \hat{\mathcal{X}}$;
  $\hat{\mathcal{X}} \leftarrow [\![\boldsymbol{\sigma}; U, V, W]\!]$;
until $\|\mathcal{X} - \hat{\hat{\mathcal{X}}}\|_F - \|\mathcal{X} - \hat{\mathcal{X}}\|_F < \epsilon$
Return $\hat{\mathcal{X}}$.

defined by

$$A \circledast B = [\boldsymbol{a_1} \otimes \boldsymbol{b_1}, \ldots, \boldsymbol{a_r} \otimes \boldsymbol{b_r}],$$

where $\otimes$ is the usual Kronecker product. An alternative description using the vectorization operator vec is:

$$A \circledast B = [\mathrm{vec}(\boldsymbol{b_1}\boldsymbol{a_1}^T), \ldots, \mathrm{vec}(\boldsymbol{b_r}\boldsymbol{a_r}^T)].$$

Finally, $A \odot B$ is the Hadamard (elementwise) product of A and B.

The ALS method is the most commonly used algorithm to compute rank-k approximations of tensors. The convergence of this algorithm can be slow in practice, and the fixed point found by the algorithm can be a nonglobal minimum or even a non-extremal stationary point of Problem (11.2).

Case of the best rank-one approximation

Even the case of the best rank-one approximation is hard. It is easy to verify that the optimization problem

$$\min_{\lambda, \boldsymbol{u} \in \mathbb{R}^{d_1}, \boldsymbol{v} \in \mathbb{R}^{d_2}, \boldsymbol{w} \in \mathbb{R}^{d_3}} \quad \|\mathcal{X} - \sigma\ \boldsymbol{u} \circ \boldsymbol{v} \circ \boldsymbol{w}\|_F \qquad \text{s.t.} \quad \|\boldsymbol{u}\| = \|\boldsymbol{v}\| = \|\boldsymbol{w}\| = 1$$

is equivalent to

$$\max_{\boldsymbol{u}\in\mathbb{R}^{d_1},\boldsymbol{v}\in\mathbb{R}^{d_2},\boldsymbol{w}\in\mathbb{R}^{d_3}} \sum_{i\in[d_1],j\in[d_2],k\in[d_3]} x_{i,j,k}u_iv_jw_k \qquad \text{s.t.} \quad \|\boldsymbol{u}\| = \|\boldsymbol{v}\| = \|\boldsymbol{w}\| = 1 \tag{11.3}$$

and the optimal value of σ is $\sum_{i,j,k} x_{i,j,k}u_iv_jw_k$ (cf. Zhang and Golub [ZG01]).

Recall that $\mathcal{X}$ defines a multilinear application mapping $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \mathbb{R}^{d_3}$ to $\mathbb{R}$. Let us denote by $\mathcal{S}_p^{d-1} = \{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_p = 1\}$ the unit sphere of $\mathbb{R}^d$ in the ℓ_p norm. We see from the equivalent formulation (11.3) that the vectors $\boldsymbol{u}, \boldsymbol{v}$ and $\boldsymbol{w}$ of the best rank-one approximation $\lambda\ \boldsymbol{u} \circ \boldsymbol{v} \circ \boldsymbol{w}$ are critical points of the multilinear form $\mathcal{X}$ restricted to $\mathcal{S}_2^{d_1-1} \times \mathcal{S}_2^{d_2-1} \times \mathcal{S}_2^{d_3-1}$. We show in Example 11.2.1 and Figure 11.9 that many such critical points can exist, and that the ALS procedure is likely to converge to a sub-optimal point.

In fact, Friedland, Gaubert and Han [FGH10] have proved an analog of Perron-Frobenius theorem for tensors, which can be stated as: *if $\mathcal{X}$ has only nonnegative entries and is* indecomposable*, and if p_1, p_2 and p_3 are larger than or equal to 3, then the multilinear form $\mathcal{X}$ restricted to $\mathcal{S}_{p_1}^{d_1-1} \times \mathcal{S}_{p_2}^{d_2-1} \times \mathcal{S}_{p_3}^{d_3-1}$ has only one critical point, which is necessary positive.* The tensor considered in Example 11.2.1 was inspired by an example of [FGH10], which illustrates that the latter results does not hold for $p_1 = p_2 = p_3 = 2$. This theorem somehow shows that the natural geometry of third-order tensors lies in ℓ_3, which explains the difficulty of minimizing the Frobenius norm.

Tucker decomposition

Another kind of decomposition was proposed by Tucker [Tuc66], in which more interaction between the different factors is allowed. The idea is to multiply a *core tensor* of dimension $r_1 \times r_2 \times r_3$ by a different matrix in each mode to form a *rank-(r_1, r_2, r_3) approximation* of $\mathcal{X}$:

$$\mathcal{X} \approx [\![\mathcal{C};\ U, V, W]\!] := \sum_{\substack{k_1\in[r_1]\\ k_2\in[r_2]\\ k_3\in[r_3]}} c_{k_1,k_2,k_3}\ \boldsymbol{u_{k_1}} \circ \boldsymbol{v_{k_2}} \circ \boldsymbol{w_{k_3}}. \tag{11.4}$$

Contrarily to the CP decomposition, is is easy to find the number of required columns R_1, R_2, and R_3 of U, V, and W, so that the decomposition (11.4) is exact: if R_i is the rank

Example 11.2.1. Consider the $2 \times 2 \times 2$ supersymmetric tensor $\mathcal{X}$ whose frontal slices are:

$$X_1 = \begin{pmatrix} 1.2 & 0.2 \\ 0.2 & 0.1 \end{pmatrix} \quad \text{and} \quad X_2 = \begin{pmatrix} 0.2 & 0.1 \\ 0.1 & 1.2 \end{pmatrix}.$$

The best rank-1 approximation problem for this tensor has two stationary points: $\mathcal{X}_1 = \sigma_1 \boldsymbol{v_1} \circ \boldsymbol{v_1} \circ \boldsymbol{v_1}$, where $\sigma_1 \approx 1.2211$ and $\boldsymbol{v_1} \approx [0.1518, 0.9884]^T$ (global minimum), and $\mathcal{X}_2 = \sigma_2 \boldsymbol{v_2} \circ \boldsymbol{v_2} \circ \boldsymbol{v_2}$, where $\sigma_2 \approx 1.2669$ and $\boldsymbol{v_2} \approx [0.9707, 0.2402]^T$ (local optimum). We noticed that the ALS procedure with a random initialization converged 37% of the time to the global optimum, and 63% of the time to the other local minimum. The (squared) objective function $\|\mathcal{X} - \hat{\mathcal{X}}\|_F^2$, as restrained to supersymmetric rank-one approximations $\hat{\mathcal{X}} = \boldsymbol{v} \circ \boldsymbol{v} \circ \boldsymbol{v}$, is plotted on Figure 11.9.

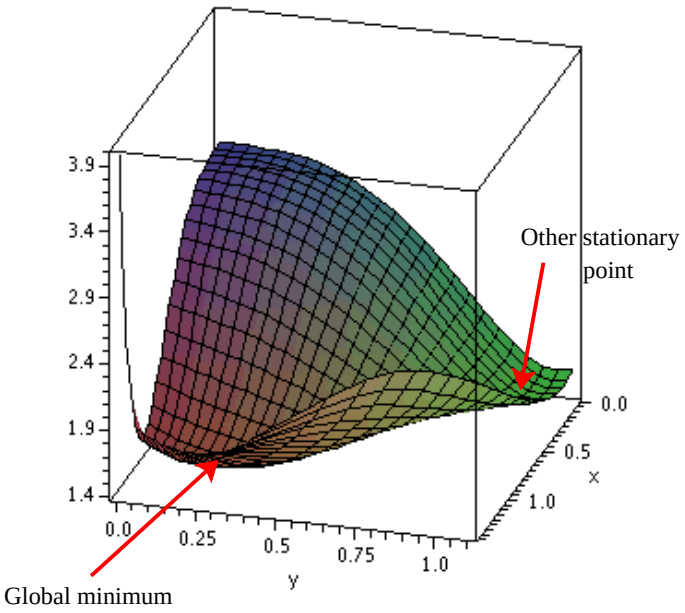


Figure 11.9: Squared objective function $\|\mathcal{X} - \hat{\mathcal{X}}\|_F^2$ of Problem (11.2), for the tensor $\mathcal{X}$ of Example 11.2.1, and supersymmetric rank-one approximations $\hat{\mathcal{X}} = \boldsymbol{v} \circ \boldsymbol{v} \circ \boldsymbol{v}$, where $\boldsymbol{v} = [x, y]^T$.

of $X_{(i)}$, and U_i contains the R_i leading left singular vectors of the $i^{\text{th}}-$mode unfolding $X_{(i)}$ (for $i = 1, 2, 3$), then

$$\mathcal{X} = \Big[\!\Big[\underbrace{[\![\mathcal{X};\ U_1^T, U_2^T, U_3^T]\!]}_{\mathcal{C}};\ U_1, U_2, U_3 \Big]\!\Big],$$

is an exact rank-(R_1, R_2, R_3) decomposition of $\mathcal{X}$. Moreover, any approximation in the form of Equation (11.4) with $r_i < R_i$ for at least one index $i \in [3]$ is necessary inexact [Tuc66].

Truncating the latter exact decomposition to the first r_i columns of U_i (for $i = 1, 2, 3$) does not yield the best rank-(r_1, r_2, r_3) approximation of $\mathcal{X}$. However, this truncation – which is now called the Higher Order SVD (HOSVD) of $\mathcal{X}$ after the work of De Lathauwer, De Moor and Vandewalle [LMV00]– can be used as a starting point for an ALS algorithm.

Another advantage of decompositions in the form of Equation (11.4) is that the columns of U, V and W, can be chosen orthonormal, which makes the interpretation easier. Let $U = Q_U R_U$, $V = Q_V R_V$ and $W = Q_W R_W$ be the QR decompositions of U, V and W. The reader can verify that

$$[\![\mathcal{C};\ U, V, W]\!] = \Big[\!\Big[[\![\mathcal{C};\ R_U, R_V, R_W]\!];\ Q_U, Q_V, Q_W \Big]\!\Big],$$

and the columns of the matrices Q_U, Q_V and Q_W in the latter decomposition are orthogonal.

Finally, we point out that alternative approaches have been recently proposed to gen-

eralize the spectral theory of matrices to higher order tensors. In [WvB10], Weiland and Van Belzen have defined a singular value decomposition for tensors which generalizes the Courant-Fisher min-max characterization of the eigenvalues of a matrix. Contrarily to most previous approaches, the tensor is seen as a multilinear form, and the proposed SVD does not depend on the set of basis for $\mathbb{R}^{d_1}$, $\mathbb{R}^{d_2}$ and $\mathbb{R}^{d_3}$ in which it is represented. In [GER10], Gnang, Elgammal and Retakh have handled the problem by mean of a ternary operator generalizing the matrix multiplication. They proved a spectral theorem which shows that every supersymmetric tensor can be decomposed as a ternary product involving a diagonal and an orthogonal tensor.

Nonnegative tensor factorization

In many applications, the multi-dimensional data contained in $\mathcal{X}$ is nonnegative. It has therefore been proposed to search for decompositions in which the different components are nonnegative. One reason is that often, such nonnegative factorizations may have a physical interpretation. In our Internet traffic problem for instance, a nonnegative decomposition of the traffic tensor would maybe reveal that the traffic is the sum of several components, accounting for different type of usage of the network. A recent book of Cichocki, Zdunek, Phan and Amari [CZPA09] review many algorithms for nonnegative matrix and tensor factorizations, as well as some applications.

The *workhorse* algorithm for tensor decomposition problems with nonnegativity constraints on the factors U, V, and W is in fact a slight modification of the ALS procedure 11.2.1 in which, after each step, the negative components are rounded to 0. This heuristic algorithm has the nice property of returning *sparse* decompositions, which is desired in many applications. Another approach presented in [CZPA09] aims at minimizing the $\alpha-$divergence $D_\alpha(\mathcal{X}\|\hat{\mathcal{X}})$, where $\hat{\mathcal{X}}$ is a rank-r tensor with nonnegative factors U,V and W. The $\alpha-$divergence between two vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ is defined as:

$$D_\alpha(\boldsymbol{p}\|\boldsymbol{q}) := \frac{1}{\alpha(\alpha-1)}\sum_i (p_i^\alpha q_i^{1-\alpha} - \alpha_i + (\alpha-1)q_i).$$

As a limiting case when $\alpha \to 1$, one obtains the (generalized) Kullback-Leibler divergence:

$$\lim_{\alpha\to 1} D_\alpha(\boldsymbol{p}\|\boldsymbol{q}) = D_{KL}(\boldsymbol{p}\|\boldsymbol{q}) = \sum_i (p_i \ln \frac{p_i}{q_i} - p_i + q_i),$$

while the dual Kullback Leibler divergence $D_{KL}(\boldsymbol{q}\|\boldsymbol{p})$ is obtained as $p \to 0$. The class of α divergences can thus be seen as smooth deformations from the Kullback Leibler divergence $D_{KL}(\boldsymbol{p}\|\boldsymbol{q})$ to its dual $D_{KL}(\boldsymbol{q}\|\boldsymbol{p})$.

One advantage to work with Kullback-Leibler (or $\alpha-$) divergences is that algorithms relying on this *metric* preserve the nonnegativity of the factors U, V, and W. Moreover, we have seen in Chapter 9 that entropy minimization problems are well-suited for traffic matrices, in particular for their information theoretic foundations. Cichocki et. al. [CZPA09]

proposed some multiplicative rules to update cyclically U, V, and W, in order to approximate the best rank-r tensor decomposition of $\mathcal{X}$ (for $D_\alpha-$divergences). The latter algorithm is in fact an *exponentiated gradient descent*.

11.2.2 Decomposition of traffic tensors

We will now study some decomposition of real three-way data. We will use the data that was already presented in Section 10.5.1, namely real traffic matrices from the Abilene network [Abi], as well as traffic matrices from the Opentransit network. We only dispose of partial measurements on Opentransit; missing values were therefore simulated by following the recommendations of Nucci, Sridharan and Taft [NST05]. Moreover, our Opentransit data consists in 40 hours of measurements only. We have extended the data to one week (168 hours) by considering the 4 leading terms of the Fourier expansion of the flows, and adding a noise. The traffic tensors $\mathcal{X}$ which we consider are thus of dimension $11 \times 11 \times 168$ for Abilene, and $116 \times 116 \times 168$ for Opentransit.

We have computed several decompositions of the traffic tensor of Abilene and Opentransit. The results are indicated in Table 11.1: for each network, we have indicated the relative L_2-error that can be obtained with different low rank tensor approximations, as well as the compression rates. Consider for example the Tucker$(6, 6, 4)-$ approximation of $\mathcal{X}$, on Abilene. This decomposition comprises a 11×6 matrix U, a 11×6 matrix V, a 168×4 matrix W, and a $6 \times 6 \times 4-$core tensor $\mathcal{C}$. The number of entries required to define the tensor $\hat{\mathcal{X}}$ is thus $11 \cdot 6 + 11 \cdot 6 + 168 \cdot 4 + 6 \cdot 6 \cdot 4 = 948$, whereas the original tensor has $11 \cdot 11 \cdot 168 = 20328$ entries. The compression rate is thus $948/20328 = 4,66\%$. This example shows that reducing the number of unknowns by a factor 20 can still yield an approximation for which the relative L_2-error $\frac{\|\mathcal{X}-\hat{\mathcal{X}}\|_F}{\|\mathcal{X}\|_F}$ is in the order of 10%. The potential of low rank tensor approximations is even more striking on Opentransit, where a relative error of only 1% can be achieved by reducing the number of variables by 100 (with the Tucker$(20, 20, 30)$-decomposition.

A general observation is that Tucker decompositions seem to yield more accurate estimations than CP decompositions for the same compression rate, in particular when the modal rank corresponding to time is small. (For a compression rate of 4.66%, we obtain a relative error of 16.5% for the CP decomposition of rank 5, vs. an error of only 12.8% for a Tucker decomposition of rank (6,6,4). Low rank decompositions based on the Kullback-Leibler divergence yield less accurate estimates (which is natural since we use a different metric than the one which is minimized, and factors are constrained to be positive), and numerical problems occurred on large tensors.

Now, we may wonder how to interpret those low rank decompositions. What is the significance of a CP (resp. Tucker) decomposition for a snapshot X_t of the traffic matrix ?

Network	Abilene		Opentransit	
Dimension of $\mathcal{X}$	$11 \times 11 \times 168$		$116 \times 116 \times 168$	
Decomposition	Fit	Compression	Fit	Compression
CP, $r = 1$	0.7134	0.93%	0.1562	0.17‰
CP, $r = 3$	0.2540	2.80%	0.0792	0.53‰
CP, $r = 5$	0.1651	4.67%	0.0629	0.88‰
CP, $r = 8$	0.1245	7.74%	0.0421	1.41‰
CP, $r = 11$	0.1053	10.28%	0.0363	1.94‰
CP, $r = 15$	0.0865	14.02%	0.0292	2.65‰
CP, $r = 30$	0.0544	28.04%	0.0203	5.30‰
CP, $r = 60$	0.0290	56.08%	0.0128	1.06%
KL, $r = 1$	0.7300	0.93%	0.4493	0.17‰
KL, $r = 3$	0.3521	2.80%	0.3221	0.53‰
KL, $r = 5$	0.2027	4.67%	Numerical issues	
KL, $r = 10$	0.1271	9.34%	Numerical issues	
Tucker, $(r_1, r_2, r_3) = (3, 3, 3)$	0.2514	2.93%	0.0729	0.54‰
Tucker, $(r_1, r_2, r_3) = (3, 3, 5)$	0.2456	4.67%	0.0665	0.69‰
Tucker, $(r_1, r_2, r_3) = (5, 5, 3)$	0.1553	3.38%	0.0620	0.76‰
Tucker, $(r_1, r_2, r_3) = (5, 5, 5)$	0.1376	5.28%	0.0482	0.94‰
Tucker, $(r_1, r_2, r_3) = (5, 5, 8)$	0.1276	8.13%	0.0429	1.19‰
Tucker, $(r_1, r_2, r_3) = (6, 6, 4)$	0.1289	4.66%	0.0515	0.97‰
Tucker, $(r_1, r_2, r_3) = (6, 6, 10)$	0.1034	10.68%	0.0352	1.51‰
Tucker, $(r_1, r_2, r_3) = (8, 8, 5)$	0.1092	6.57%	0.0409	1.33‰
Tucker, $(r_1, r_2, r_3) = (8, 8, 12)$	0.0752	14.56%	0.0291	2.05‰
Tucker, $(r_1, r_2, r_3) = (9, 9, 6)$	0.0993	8.32%	0.0359	1.58‰
Tucker, $(r_1, r_2, r_3) = (9, 9, 9)$	0.0789	11.99%	0.0305	1.91‰
Tucker, $(r_1, r_2, r_3) = (15, 15, 5)$		-	0.0375	2.40‰
Tucker, $(r_1, r_2, r_3) = (15, 15, 15)$		-	0.0207	4.14‰
Tucker, $(r_1, r_2, r_3) = (20, 20, 15)$		-	0.0191	5.82‰
Tucker, $(r_1, r_2, r_3) = (20, 20, 30)$		-	0.0135	9.59‰
Tucker, $(r_1, r_2, r_3) = (30, 30, 20)$		-	0.0147	1.25%

Table 11.1: Fit (L2 relative error $\frac{\|\mathcal{X}-\hat{\mathcal{X}}\|_F}{\|\mathcal{X}\|_F}$) and Compression rates for several decompositions $\hat{\mathcal{X}}$ of the traffic tensor $\mathcal{X}$.

For a CP decomposition $\hat{\mathcal{X}} = [\![U, V, W]\!]$ of rank r, the elements of a slice of $\hat{\mathcal{X}}$ are

$$(\hat{X}_t)_{o,d} = \hat{x}_{o,d,t} = \sum_{k=1}^{r} U_{o,k} V_{d,k} W_{t,k} = \sum_{k=1}^{r} W_{t,k} (\boldsymbol{u_k} \boldsymbol{v_k}^T)_{o,d}.$$

Thus, each snapshot of the TM is approximated by a linear combination of the rank-one matrices $\boldsymbol{u_k} \boldsymbol{v_k}^T$ (for $k \in [r]$). For a Tucker decomposition $\hat{\mathcal{X}} = [\![\mathcal{C}; U, V, W]\!]$ of rank (r_1, r_2, r_3),

$$(\hat{X}_t)_{o,d} = \hat{x}_{o,d,t} = \sum_{k_1=1}^{r_1} \sum_{k_2=1}^{r_2} \sum_{k_1=1}^{r_3} c_{k_1,k_2,k_3} \; U_{o,k_1} V_{d,k_2} W_{t,k_3} = \sum_{k_3=1}^{r_3} W_{t,k_3} (U C^{(k_3)} V^T)_{o,d},$$

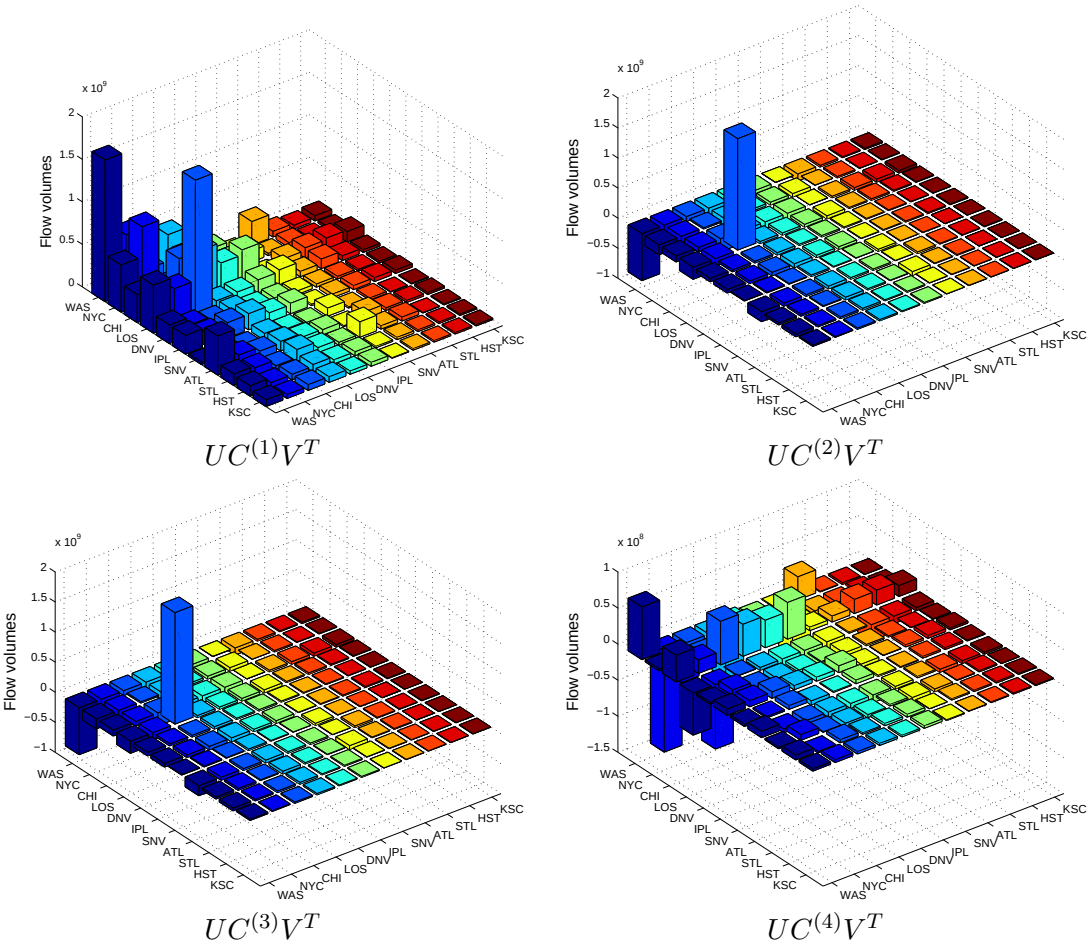


Figure 11.10: Four principal components for the temporal slices (snapshots) of the Abilene traffic tensor.

where $C^{(k_3)}$ is the slice of $\mathcal{C}$ with k_3 fixed in the third mode: $C^{(k_3)}_{k_1,k_2} = c_{k_1,k_2,k_3}$. This means that the snapshots of $\hat{\mathcal{X}}$ are linear combinations of the matrices $UC^{(k_3)}V^T$ (for $k_3 \in [r_3]$). We have plotted the entries of the matrices $UC^{(1)}V^T, \ldots, UC^{(4)}V^T$, for the Tucker(6,6,4) approximation of the Abilene traffic tensor on Figure 11.10. For the observation period considered, each TM can be decomposed as a linear combination of these 4 matrices with a relative L_2-error below 10%.

11.2.3 Using tensor decompositions for the estimation of Traffic matrices

We propose in this section the preliminary sketch of a method relying on tensor to estimate traffic matrices. Based on the observations from the previous section, we propose to compute, at time step t, the tensor decomposition (CP or Tucker) of the estimate traffic tensor $\hat{\mathcal{X}}_{t-\delta:t-1}$ on the sliding window from time $t-\delta$ to $t-1$, and to use this decomposition to produce basis matrices $B_1, \ldots, B_r$ (we take $B_i = \boldsymbol{u_i}\boldsymbol{v_i}^T$ for CP decomposition $\hat{\mathcal{X}}_{t-\delta:t-1} = [\![U, V, W]\!]$ or $B_i = UC^{(i)}V^T$ for a Tucker decomposition $\hat{\mathcal{X}}_{t-\delta:t-1} = [\![\mathcal{C}; U, V, W]\!]$). Then, we

estimate the snapshot of the traffic matrix at time t by searching the weights $w_1^{(t)}, \ldots, w_r^{(t)}$ which minimize the L_2-error on the measurement equation:

$$\left\| A \operatorname{vec} \left(\sum_{k=1}^{r} w_k^{(t)} B_k \right) - \boldsymbol{y}_t \right\|.$$

Finally, we apply the IPF algorithm. We have written the pseudo code of this technique for Tucker decompositions, called $T4$: *Tucker Traffic Tensor Tomography*, in Algorithm 11.2.2. Note that this method can be used online.

Algorithm 11.2.2 Tucker Traffic Tensor Tomography (T4)

Input: Rank (r_1, r_2, r_3) of the Tucker decompositions
Input: width δ of the sliding window
for $t = 1, \ldots, \delta$ **do**
  Read the measurement $\boldsymbol{y_t}$, and compute the tomogravity estimate $\hat{X}_t$ of the TM at time t;
end for
for $t = \delta + 1, \ldots, T$ **do**
  Form the tensor $\hat{\mathcal{X}}_{t-\delta:t-1}$ whose slices are the δ previously estimated snapshots

$$\hat{X}_{t-\delta}, \ldots, \hat{X}_{t-1};$$

  Compute a Tucker approximation $[\![\mathcal{C}; U, V, W]\!]$ of $\hat{\mathcal{X}}_{t-\delta:t-1}$;
  Read the measurement $\boldsymbol{y_t}$, and compute the weights $\boldsymbol{w^{(t)}}$ of the new estimate:

$$\boldsymbol{w^{(t)}} \leftarrow \left(A(V \otimes U) C_{(3)}^T \right)^{\dagger} \boldsymbol{y_t};$$

  Compute the prior estimate $\tilde{X}_t = \sum_{k=1}^{r_3} w_k^{(t)} U C^{(k)} V^T$;
  Compute the estimate $\hat{X}_t$ with the IPF algorithm, by using the prior $\tilde{X}_t$;
end for

We have used our $T4$ algorithm to compute an estimate of the Opentransit traffic tensor. We have used rank-$(30, 30, 20)$ Tucker decompositions, and a time window of length $\delta = 24$. We rely on SNMP measurements only. The temporal L_2-error is plotted on Figure 11.11, and compared with the error of the raw gravity estimate and the tomogravity estimate. The tomogravity estimate is the same as our $T4$ estimate during the first 24 hours. Then, when we start to use tensor decompositions, the L_2-error jumps from roughly 30% to 20%. This looks very promising, since the use of tensors yields an improvement which is comparable to the improvement observed when entropic projections (IPF) of the raw gravity model are done (the error jumps from roughly 40 to 30%). We must nevertheless express some reservations about the data used for this experiment: recall that some parts of our Opentransit data was synthetically generated. There is a risk that our results might not be reproducible with real data. For future work, we would like to evaluate our $T4$ approach in presence of Netflow measurements, and validate it with real data from a large network.

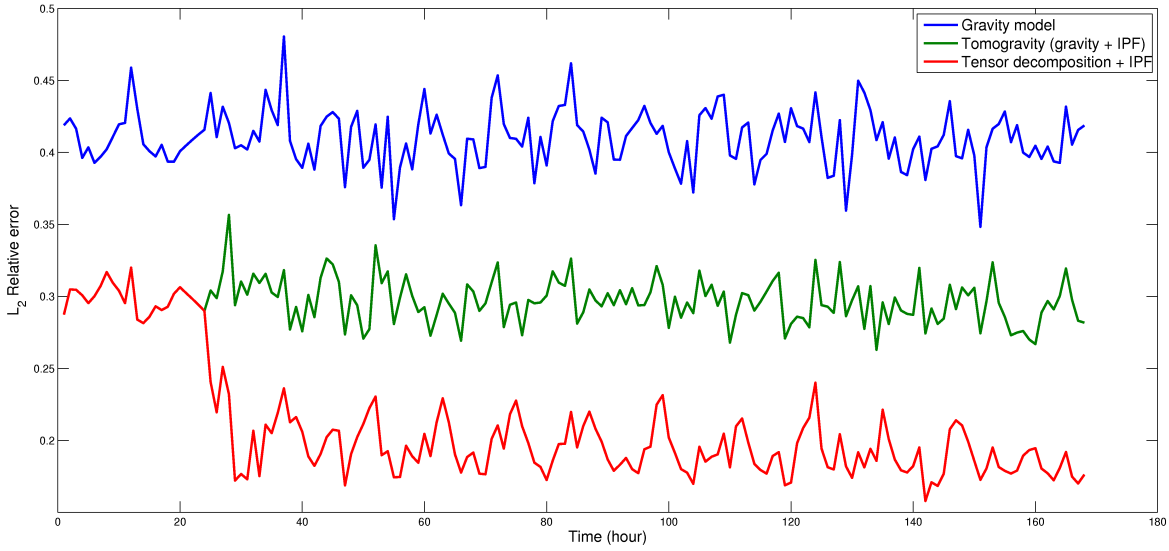


Figure 11.11: Temporal error on Opentransit, for three estimates of the flows based on link counts only (gravity, tomogravity, and T4).

Bibliography

[AB01] A.C. Atkinson and R.A. Bailey. One hundred years of the design of experiments on and off the pages of Biometrika. *Biometrika*, 88(1):53–97, 2001.

[Abi] Abilene measurements: `http://www.cs.utexas.edu/~yzhang/research/Abilene-TM/`.

[AGLN06] Marianne Akian, Stephane Gaubert, Bas Lemmens, and Roger Nussbaum. Iteration of order preserving subhomogeneous maps on a cone. *Mathematical Proceedings of the Cambridge Philosophical Society*, 140:157, 2006.

[AO82] T . Abatzoglou and B. O'Donnell. Minimization by coordinate descent. *Journal of Optimization Theory and Applications*, 36:163–174, 1982.

[ARV09] S. Arora, S. Rao, and U. Vazirani. Expander flows, geometric embeddings and graph partitioningc. *Journal of the ACM*, 56(2):1–37, 2009.

[AS04] A. A. Ageev and M. I. Sviridenko. Pipage rounding: a new method of constructing algorithms with proven performance guarantee. *Journal of Combinatorial Optimization*, 8(3):307–328, 2004.

[AS08] A. Babapour Atashgah and A. Seifi. Optimal design of multi-response experiments using semi-definite programming. *Journal of Optimization and Engineering*, 10:75–90, 2008.

[Atw73] C.L. Atwood. Sequences converging to D-optimal designs of experiments. *Annals of statistics*, 1(2):342–352, 1973.

[Atw76] C. L. Atwood. Convergent design sequences, for sufficiently regular optimality criteria. *The Annals of Statistics*, 4(6):1124–1138, 1976.

[Atw80] C. L. Atwood. Convergent design sequences for sufficiently regular optimality criteria II. Singular case. *The Annals of Statistics*, 8(4):894–912, 1980.

[AZ99] T. Ando and X. Zhan. Norm inequalities related to operator monotone functions. *Mathematische Annalen*, 315:771–780, 1999.

[Bar95] A. I. Barvinok. Problems of distance geometry and convex properties of quadratic maps. *Discrete and computational Geometry*, 13:189–202, 1995.

[BB96] H. Bauschke and J. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM Review*, 38(3):367–426, 1996.

[BBL95] H.H. Bauschke, J.M. Borwein, and A.S. Lewis. The method of cyclic projections for closed convex sets in Hilbert space. *Contemporary Mathematics*, 204:1–38, 1995.

[Ber95] D.P. Bertsekas. *Nonlinear programming*. Belmont, MA: Athena Scientific, 1995.

[BGS08] M. Bouhtou, S. Gaubert, and G. Sagnol. Optimization of network traffic measurement: a semidefinite programming approach. In *Proceedings of the International*

Conference on Engineering Optimization (ENGOPT), Rio De Janeiro, Brazil, 2008. ISBN 978-85-7650-152-7.

[BGS10] M. Bouhtou, S. Gaubert, and G. Sagnol. Submodularity and randomized rounding techniques for optimal experimental design. *Electronic Notes in Discrete Mathematics*, 36:679 – 686, March 2010. ISCO 2010 - International Symposium on Combinatorial Optimization. Hammamet, Tunisia.

[BK07] M. Bouhtou and O. Klopfenstein. Robust optimization for selecting Netflow points of measurement in an ip network. In *GLOBECOM '07, IEEE*, pages 2581–2585, 2007.

[BM03] S. Burer and R.D.C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming (series B)*, 95(2):329–357, 2003.

[BM05] S. Burer and R.D.C. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming (series A)*, 103(3):427–444, 2005.

[BR97] R.B. Bapat and T.E.S. Raghavan. *Nonnegative Matrices and Applications*. Cambridge University Press, 1997.

[BR04] N. Benameur and J. Roberts. Traffic matrix inference in IP networks. *Networks and Spatial economics*, 4(1):7–21, 2004. Special Issue on "crossovers between transportation planning and telecommunications".

[Bre67a] L.M. Bregman. Proof of the convergence of Sheleikhovskii's method for a problem with transportation constraints. *USSR Computational Mathematics and Mathematical Physics*, 7(1):191–204, 1967.

[Bre67b] L.M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

[Bru68] R.A. Brualdi. Convex sets of nonnegative matrices. *Canadian Journal of Mathematics*, 29:144–157, 1968.

[BTMZ90] A. Ben-Tal, A. Melman, and J. Zowe. Curved search methods for unconstrained optimization. *Optimization*, 21:669–695, 1990.

[BTN92] A. Ben-Tal and A. Nemirovskii. Interior point polynomial-time method for truss topology design. Technical report, Faculty of Industrial Engineering and Management, Technion institute of Technology, Haifa, Israel, 1992.

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[BVJ06] P. Bermolen, S. Vaton, and I. Juva. Search for optimality in traffic matrix estimation: a rational approach by Cramer-Rao lower bounds. In *NGI'06 : 2nd Conference on Next Generation Internet Design and Engineering, Valencia*, pages 224–231, 2006.

[BWKR08] M. Branderhorst, I. Walmsley, R. Kosut, and H. Rabitz. Optimal experiment design for quantum state tomography of a molecular vibrational mode. *Journal of Physics B: Atomic Molecular and Optical Physics*, 41:074004, 2008.

[CB05] B.Y. Choi and S. Bhattacharyya. On the accuracy and overhead of Cisco sampled Netflow. In *ACM SIGMETRICS Workshop on Large Scale Network Inference (LSNI)*, Banff, Canada, June 2005.

[CC70] J.D. Carroll and J.J. Chang. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35:283–319, 1970.

[CC84] M. Conforti and G. Cornuéjols. Submodular set functions, matroids and the greedy algorithm: tight worst-case bounds and some generalizations of the Rado-Edmonds theorem. *Discrete applied mathematics*, 7(3):251–274, 1984.

[CCPa07] G. Calinescu, C. Chekuri, M. Pál, and J. Vondr /ák. Maximizing a submodular set function subject to a matroid constraint. In *Proceedings of the 12th international conference on Integer Programming and Combinatorial Optimization, IPCO*, volume 4513, pages 182–196, 2007.

[CDPE+90] Y. Censor, A.R. De Pierro, T. Elfving, G.T. Herman, and A.N. Iusem. On iterative methods for linearly constrained entropy maximization. *Banach center publications*, 24:145–163, 1990.

[CDVY00] J. Cao, D. Davis, S. Vander Wiel, and B. Yu. Time-varying network tomography: Router link data. *Journal of the American Statistical Association*, 95:1063–1075, 2000.

[CF95] D. Cook and V. Fedorov. Constrained optimization of experimental design. *Statistics*, 26:129–178, 1995.

[Che99] H. Chernoff. Gustav Elfving's impact on experimental design. *Statistical Science*, 14(2):201–205, 1999.

[CIB+06] G.R. Cantieni, G. Iannaccone, C. Barakat, C. Diot, and P. Thiran. Reformulating the monitor placement problem: Optimal network-wide sampling. In *Proceedings of the 2006 ACM CoNEXT conference*, pages 1–12. ACM, 2006.

[CISa] CISCO. Netflow services solutions guide. `http://www.cisco.com/en/US/docs/ios/solutions_docs/netflow/nfwhite.html`.

[CISb] CISCO. Netflow white papers. `http://www.cisco.com/en/US/products/ps6601/prod_white_papers_list.html`.

[CIS07] CISCO. Netflow performance analysis, technical white paper, May 2007.

[CVFC09] P. Casas, S. Vaton, L. Fillatre, and T. Chonavel. Efficient methods for traffic matrix modeling and on-line estimation in large-scale ip networks. In *21st International teletraffic congress ITC 21*, Paris, France, September 2009.

[CZPA09] A. Cichocki, R. Zdunek, A.H. Phan, and S. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation; electronic version*. Wiley, 2009.

[dAEJL07] A. d' Aspremont, L. El Ghaoui, M.I. Jordan, and G.R.G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3):434–448 (electronic), 2007.

[Det93] H. Dette. Elfving's theorem for D-optimality. *The Annals of Statistics*, 21:753–766, 1993.

[DH94] F. Deutsch and H. Hundal. The rate of convergence of Dykstra's cyclic projections algorithm: the polyhedral case. *Numerical Functional Analysis and Optimization*, 15(5–6):537–556, 1994.

[DHL09] H. Dette and T. Holland-Letz. A geometric characterization of c-optimal designs for heteroscedastic regression. *The Annals of Statistics*, 37(6B):4088–4103, December 2009.

[DLR77] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood for incomplete data via the EM algorithm, with discussion. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[DPZ08] H. Dette, A. Pepelyshev, and A. Zhigljavsky. Improving updating rules in multiplicative algorithms for computing D-optimal designs. *Computational Statistics & Data Analysis*, 53(2):312 – 320, 2008.

[DS93] H. Dette and W.J. Studden. Geometry of E-optimality. *The Annals of Statistics*, 21:416–433, 1993.

[Ede89] A. Edelman. *Eigenvalues and condition numbers of random matrices*. PhD thesis, Department of Mathemathics, Massachusetts Institute of Technology, Cambridge, MA, USA, 1989.

[Elf52] G. Elfving. Optimum allocation in linear regression theory. *The Annals of Mathematical Statistics*, 23:255–262, 1952.

[Elf80] T. Elfving. On some methods for entropy maximization and matrix scaling. *Linear Algebra and its Applications*, 34:321–339, 1980.

[ER05] A. Edelman and N.R. Rao. Random matrix theory. *Acta Numerica*, 14:233–297, 2005.

[Fed72] V.V. Fedorov. *Theory of optimal experiments*. New York : Academic Press, 1972. Translated and edited by W. J. Studden and E. M. Klimko.

[Fei98] U. Feige. A threshold of $\ln n$ for approximating set cover. *Journal of ACM*, 45(4):634–652, July 1998.

[FGH10] S. Friedland, S. Gaubert, and L. Han. Perron-frobenius theorem for nonnegative multilinear forms and extensions. arXiv:0905.1626, 2010.

[FGL+01] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, J. Rexford, and F. True. Deriving traffic demands for operational IP networks: Methodology and experience. *IEEE/ACM Transactions on Networking*, 9(3):265–280, 2001.

[Fis35] R.A. Fisher. *The design of experiments*. Edinburgh: Oliver & Boyd, 1935.

[FL89] J. Franklin and J. Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114/115:717–735, 1989.

[FL00] V. Fedorov and J. Lee. Design of experiments in statistics. In H.Wolkowicz, R.Saigal, and L.Vandenberghe, editors, *Handbook of semidefinite programming*, chapter 17. Kluwer, 2000.

[FRT97] S.C. Fang, J.R. Rajasekera, and H.-S.J. Tsao. *Entropy Optimization and Mathematical Programming*. Kluwer, Norwell, MA, 1997.

[GBH70] R. Gordon, R. Bender, and G.T. Herman. Algebraic reconstruction techniques (art) for three dimensional electron microscopy and X-ray photography. *Journal of theoretical Biology*, 29:471–481, 1970.

[GER10] E.K. Gnang, A. Elgammal, and V. Retakh. A generalized spectral theory for tensors. arXiv:1008.2923, 2010.

[GW95] M.X. Goemans and S.P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problem using semidefinite programming. *Journal of the ACM*, 42(6):1115–1145, 1995.

[Har70] R.A. Harshman. Foundations of the parafac procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics*, 16:1–84, 1970.

[HHS95] H.Dette, B. Heiligers, and W.J Studden. Minimax designs in linear regression models. *The Annals of Statistics*, 23(1):30–40, 1995.

[HJ08] R. Harman and T. Jurík. Computing c-optimal experimental designs using the simplex method of linear programming. *Computational Statistics and data analysis*, 53:247–254, 2008.

[HJT07] R. Harman, T. Jurík, and M. Trnovská. Constructing optimal designs on finite experimental domains using methods of mathematical programming. In *8th International Workshop in Model-Oriented Design and Analysis MODA8*, Almagro, Spain, June 2007. Sildes : `http://areaestadistica.uclm.es/moda/moda8/html/slides/mODa8_HarTrnJur.pdf`.

[HP95] F. Hansen and G.K. Pedersen. Perturbation formulas for traces on C*-algebras. *Publications of the research institute for mathematical sciences, Kyoto University*, 31:169–178, 1995.

[HP07] R. Harman and L. Pronzato. Improvements on removing nonoptimal support points in D-optimum design algorithms. *Statistics & Probability letters*, 77:90–94, 2007.

[HT09] R. Harman and M. Trnovská. Approximate D-optimal designs of experiments on the convex hull of a finite set of information matrices. *Mathematica Slovaca*, 59(5):693–704, December 2009.

[IPS05] G. Iyengar, D. J. Phillips, and C. Stein. Approximation algorithms for semidefinite packing problems with applications to maxcut and graph coloring. In *Proceedings of the 11th conference on Integer Programming and Combinatorial Optimization*, volume 3509 of *Lecture Notes in Computer Science*, pages 152–166. Springer, 2005.

[Ius91] A.N. Iusem. On dual convergence and the rate of primal convergence of Bregman's convex programming method. *SIAM Journal on Optimization*, 1(3):401–423, Aug 1991.

[Jam64] T. James. Distributions of matrix variates and latent roots derived from normal samples. *Annals of Mathematical Statistics*, 35:475–501, 1964.

[JB06] E. Jorswieck and H. Boche. *Majorization and matrix-monotone functions in wireless communications*. Now Publishers Inc., 2006.

[KB09] T.G. Kolda and B.W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[Kie74] J. Kiefer. General equivalence theory for optimum designs (approximate theory). *The annals of Statistics*, 2(5):849–879, 1974.

[Kie75] J. Kiefer. Optimal design: Variation in structure and performance under change of criterion. *Biometrika*, 62(2):277–288, 1975.

[KMS98] D. Karger, R. Motwani, and M. Sudan. Approximate graph coloring by semidefinite programming. *Journal of the ACM*, 45(2):246–265, 1998.

[Kos06] T. Kosem. inequalities between $|f(a+b)|$ and $|f(a)+f(b)|$. *Linear Algebra and its Applications*, 418:153–160, 2006.

[Kru79] R.S. Krupp. Properties of Kruithof's projection method. *The Bell systems technical journal*, 58:517–538, 1979.

[KST09] A. Kulik, H. Shachnai, and T. Tamir. Maximizing submodular set functions subject to multiple linear constraints. In *SODA '09: Proceedings of the Nineteenth Annual ACM -SIAM Symposium on Discrete Algorithms*, pages 545–554, Philadelphia, PA, USA, 2009.

[KW60] J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

[Läu74] E. Läuter. Experimental design in a class of models. *Statistics*, 5(4–5):379–398, 1974.

[LCPB00] L. Laloux, P. Cizeau, M. Potters, and J.P. Bouchaud. Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, 3(3):391–398, 2000.

[LK94] D.M. Levinson and A. Kumar. A multi-modal trip distribution model: Structure and application. *Transportation Research Record*, 1466:124–131, 1994.

[LMV00] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278 (electronic), 2000.

[Lov79] L. Lovász. On the Shannon capacity of a graph. *IEEE Transactions on Information Theory*, 25:1–7, 1979.

[Löw34] K. Löwner. Über monotone Matrixfunktionen. *Mathematische Zeitschrift*, 38(1):177–216, 1934.

[LPC+04] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. Kolaczyk, and N. Taft. Structural analysis of network traffic flows. In *Proc. ACM SIGMETRICS*, 2004.

[LTY06] G. Liang, N. Taft, and B. Yu. A fast lightweight approach to origin-destination IP traffic estimation using partial measurements. *IEEE/ACM Trans. Netw.*, 14(SI):2634–2648, 2006.

[LVBL98] M.S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret. Applications of second-order cone programming. *Linear Algebra and its Applications*, 284:193–228, 1998.

[Men67] M. V. Menon. Reduction of a matrix with positive elements to a doubly stochastic matrix. *Proceedings of the American Mathematical Society*, 18(2):244–247, 1967.

[Min78] M. Minoux. Accelerated greedy algorithms for maximizing submodular set functions. In J. Stoer, editor, *Optimization Techniques*, volume 7 of *Lecture Notes in Control and Information Sciences*, pages 234–243. Springer Berlin / Heidelberg, 1978.

[Mit74] T.J. Mitchell. An algorithm for the construction of "D-optimal" experimental designs. *Technometrics. A Journal of Statistics for the Physical, Chemical and Engineering Sciences*, 16:203–210, 1974.

[MS69] M. V. Menon and H. Schneider. The spectrum of a nonlinear operator associated with a matrix. *Linear Algebra and its Applications*, 2:321–334, 1969.

[MTS+02] A. Medina, N. Taft, K. Salamatian, S. Battacharya, and C. Diot. Traffic matrix estimation: Existing techniques and new directions. In *Proc. of SIGCOMM*, pages 161–174, Pittsburgh, August 2002.

[NN94] Y. Nesterov and A. Nemirovsky. *Interior-Point polynomial algorithms in convex programming*, volume 13 of *SIAM Studies in Applied Mathematics*. SIAM, 1994.

[NRT99] A. Nemirovski, C. Roos, and T. Terlaky. On maximization of quadratic form over intersection of ellipsoids with common center. *Mathematical programming*, 86:463–473, 1999.

[NST05] A. Nucci, A. Sridharan, and N. Taft. The problem of synthetically generating ip traffic matrices: initial recommendations. *ACM SIGCOMM Computer Communication Review*, 35(3):19–32, 2005.

[NWF78] G.L. Nemhauser, L.A. Wolsey, and M.L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14:265–294, 1978.

[Pat98] G. Pataki. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of Operations Research*, 23(2):339–358, 1998.

[Pro03] L. Pronzato. Removing non-optimal support points in D-optimum design algorithms. *Statistics & Probability letters*, 63:223–228, July 2003.

[PT83] F. Pukelsheim and D.M. Titterington. General differential and lagrangian theory for optimal experimental design. *The Annals of Statistics*, 11(4):1060–1068, 1983.

[PT91] F. Pukelsheim and B. Torsney. Optimal weights for experimental designs on linearly independent support points. *The annals of Statistics*, 19(3):1614–1625, 1991.

[PTL04] K. Papagiannaki, N. Taft, and A. Lakhina. A distributed approach to measure IP traffic matrices. In *IMC'04*, Taormina,Sicily,Italy, October 2004.

[Puk80] F. Pukelsheim. On linear regression designs which maximize information. *Journal of statistical planning and inferrence*, 4:339–364, 1980.

[Puk93] F. Pukelsheim. *Optimal Design of Experiments*. Wiley, 1993.

[Qi09] H.-D. Qi. A semidefinite programming study of the elfving theorem. Technical report, University of Southampton, July 2009.

[Ric08] P. Richtarik. Simultaneously solving seven optimization problems in relative scale. Optimization online, preprint number 2185, 2008.

[Roc70] R.T. Rockafellar. *Convex analysis*. Princeton University Press, Princeton, NJ, 1970.

[RS89] T.G. Robertazzi and S.C. Schwartz. An Accelerated Sequential Algorithm for Producing D-Optimal Designs. *SIAM Journal on Scientific and Statistical Computing*, 10:341, 1989.

[Sag09a] G. Sagnol. A class of semidefinite programs with a rank-one solution. Submitted. Preprint arXiv:0909.5577, 2009.

[Sag09b] G. Sagnol. Computing optimal designs of multiresponse experiments reduces to second-order cone programming. Accepted for publication in Journal of Statistical Planning and Inference. To appear. Preprint arXiv:0912.5467, 2009.

[Sag10] G. Sagnol. Polynomial-time approximability results for combinatorial problems arising in optimal experimental design. Submitted. Preprint arXiv:1007.4152, 2010.

[SBG09] G. Sagnol, M. Bouhtou, and S. Gaubert. Optimizing the measurement of the traffic in lage scale networks : An experimental design approach. In *International Network Optimization Conference, INOC'09*, April 2009. slides: `http://www.cmap.polytechnique.fr/~sagnol/papers/presInoc_Sagnol.pdf`.

[SBG10] G. Sagnol, M. Bouhtou, and S. Gaubert. Successive c-optimal designs: a scalable technique to optimize the measurements on large networks. In *Proceedings of the ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, SIGMETRICS '10, pages 347–348, New York, NY, USA, 2010.

[SGB10] G. Sagnol, S. Gaubert, and M. Bouhtou. Optimal monitoring on large networks by successive c-optimal designs. In *22nd international teletraffic congress (ITC22), Amsterdam, The Netherlands*, September 2010. Preprint: `http://www.cmap.polytechnique.fr/~sagnol/papers/ITC22_submitted.pdf`.

[SLT+05] A. Soule, A. Lakhina, N. Taft, K. Papagiannaki, K. Salamatian, A. Nucci, M. Crovella, and Ch. Diot. Traffic matrices: balancing measurements, inference and modeling. In *SIGMETRICS '05*, pages 362–373, New York, NY, USA, 2005.

[SM08] H. Singhal and G. Michailidis. Optimal sampling in state space models with applications to network monitoring. In *SIGMETRICS'08*, Annapolis, Maryland, USA, 2008.

[SNC+07] A. Soule, A. Nucci, R. Cruz, E. Leonardi, and N. Taft. Estimating dynamic traffic matrices by using viable routing changes. *IEEE/ACM Transactions on Networking (TON)*, 15(3):485–498, june 2007.

[SQZ06] H.H. Song, L. Qiu, and Y. Zhang. Netquest: A flexible framework for largescale network measurement. In *ACM SIGMETRICS'06*, St Malo, France, 2006.

[SSNT05] A. Soule, K. Salamatian, A. Nucci, and N. Taft. Traffic matrix tracking using Kalman filters. *Proceedings of the ACM SIGMETRICS Performance Evaluation Review (PER), Special issue on the First ACM Sigmetrics Workshop on Large Scale Networks Inference (LSNI)*, 33(3), December 2005.

[ST73] S.D. Silvey and D.M. Titterington. A geometric approach to optimal design theory. *Biometrika*, 60(1):21–32, 1973.

[Ste80] G. W. Stewart. The efficient generation of random orthogonal matrices with an application to condition estimators. *SIAM Journal on Numerical Analysis*, 17(3):403–409, 1980.

[STT78] S.D. Silvey, D.M. Titterington, and B. Torsney. An algorithm for optimal designs on a finite design space. *Communications in Statistics - Theory and Methods*, 7(14):1379–1389, 1978.

[Stu17] "Student". Tables for estimating the probability that the mean of a unique sample of observations lies between $-\infty$ and any given distance of the mean of the population from which the sample is drawn. *Biometrika*, 11:414–417, 1917.

[Stu71] W.J. Studden. Elfving's theorem and optimal designs for quadratic loss. *The Annals of Mathematical Statistics*, 42(5):1613–1621, 1971.

[Stu99] J.F. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11–12:625–653, 1999.

[Stu05] W.J. Studden. Elfving's theorem revisited. *Journal of statistical planning and Inference*, 130:85–94, 2005.

[Svi04] M. Sviridenko. A note on maximizing a submodular set function subject to a knapsack constraint. *Operation Research Letters*, 32(1):41–43, 2004.

[Sze94] M. Szegedy. A note on the ϑ number of Lovász and the generalized Delsarte bound. In *SFCS'94: Proceedings of the 35th Annual Symposium on Foundations of Computer Science*, pages 36–39, Washington, DC, USA, 1994. IEEE Computer Society.

[Tit76] D.M. Titterington. Algorithms for computing D-optimal design on finite design spaces. In *Proceedings of the 1976 Conf. on Information Science and Systems*, pages 213–216, Baltimore, USA, 1976. Dept. of Electronic Engineering, John Hopkins University.

[Tit78] D.M. Titterington. Estimation of correlation coefficients by ellipsoidal trimming. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 27(3):227–234, 1978.

[Tor83] B. Torsney. A moment inequality and monotonicity of an algorithm. In K.O. Kurtanek and A.V. Fiacco, editors, *Proccedings of the International Symposium on Semi-infinite programming and applications, Lecture Notes in Economics and Mathemathical Systems (215)*, pages 249–260. Springer, 1983.

[Tuc66] L.R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311, 1966.

[UP07] D. Uciński and M. Patan. D-optimal design of a monitoring network for parameter estimation of distributed systems. *Journal of Global Optimization*, 39(2):291–322, 2007.

[Var96] Y. Vardi. Network tomography: Estimating source-destination intensities from link data. *Journal of the American Statistical Association*, 91:365–377, 1996.

[VBG05] S. Vaton, J.S. Bedo, and A. Gravey. Advanced methods for the estimation of the origin destination traffic matrix. In *Performance Evaluation and Planning Methods for the Next Generation Internet*. Springer, 2005.

[VBW98] L. Vandenberghe, S. Boyd, and S. Wu. Determinant maximization with linear matrix inequality constraints. *SIAM Journal on Matrix Analysis and Applications*, 19:499–533, 1998.

[Von08] J. Vondrák. Optimal approximation for the submodular welfare problem in the value oracle model. In *ACM Symposium on Theory of Computing, STOC'08*, pages 67–74, 2008.

[Wu83] C. F. J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11(1):95—103, 1983.

[WvB10] S. Weiland and F. van Belzen. Singular value decompositions and low rank approximations of tensors. *IEEE transactions on signal processing*, 58(3):1171–1182, 2010.

[WW78] C. Wu and H.P. Wynn. The convergence of general step-length algorithms for regular optimum design criteria. *The Annals of Statistics*, 6(6):1273–1285, 1978.

[Wyn70] H.P. Wynn. The sequential generation of D-optimum experimental designs. *Annals of Mathematical Statistics*, 41:1655–1664, 1970.

[Yu10a] Y. Yu. Monotonic convergence of a general algorithm for computing optimal designs. *The Annals of Statistics*, 38(3):1593–1606, 2010.

[Yu10b] Y. Yu. Strict monotonicity and convergence rate of titterington's algorithm for computing d-optimal designs. *Computational Statistics & Data Analysis*, 54(6):1419–1425, 2010.

[ZG01] T. Zhang and G.H. Golub. Rank-one approximation to high order tensors. *SIAM Journal on Matrix Analysis and Applications*, 23(2):534–550 (electronic), 2001.

[Zha02] X. Zhan. *Matrix Inequalities (Lecture Notes in Mathematics)*. Springer, 2002.

[ZN05] H. Zang and A. Nucci. Optimal netflow deployment in IP networks. In *19th International Teletraffic Congress (ITC)*, Beijing, China, August 2005.

[ZRDG03] Y. Zhang, M. Roughan, N. Duffield, and A. Greenberg. Fast accurate computation of large-scale IP traffic matrices from link loads. In *SIGMETRICS '03*, pages 206–217, 2003.

[ZRLD05] Y. Zhang, M. Roughan, C. Lund, and D.L. Donoho. Estimating point-to-point and point-to-multipoint traffic matrices: an information-theoretic approach. *IEEE/ACM Transactions on Networking*, 13(5):947–960, 2005.

[ZRWQ09] Y. Zhang, M. Roughan, W. Willinger, and L. Qiu. Spatio-temporal compressive sensing and Internet traffic matrices. In *SIGCOMM'09*, Barcelona, Spain, August 2009.

[ZW80] H.G. Van Zuylen and L.G. Willumsen. The most likely trip matrix estimated from traffic counts. *Transportation Research*, 14(3):281–293, 1980.

Plans d'expériences optimaux et application à l'estimation des matrices de trafic dans les grands réseaux. *Programmtion conique du second ordre et sous-modularité.*

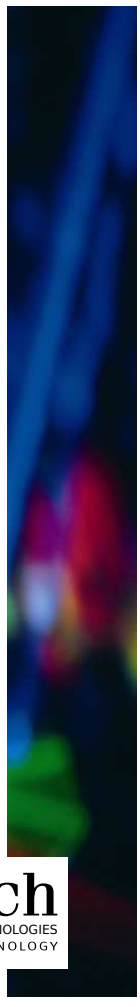
Résumé : Nous abordons le problème de l'optimisation des mesures dans les grands réseaux Internet par la théorie des plans d'expériences optimaux. Cette approche donne lieu d'étudier des problèmes de grande taille en conception optimale d'expériences, pour lesquels nous développons une méthode de résolution fondée sur l' Optimisation Conique du Second Ordre. Le coeur de notre méthode est un théorème de réduction du rang en optimisation semi-définie. Certains aspects combinatoires sont également étudiés.
L'application à l'inférence des matrices de trafic dans les réseaux IP fait l'objet de la seconde partie de ce manuscrit. Nous développons une méthode où l'on optimise l'estimation de plusieurs combinaisons linéaires (tirées de façon aléatoire) des demandes de trafic. Nous comparons notre approche aux précédentes au travers de simulations sur des données réelles. En particulier, nous traitons des instances pour lesquelles les approches précédentes étaient incapables de fournir une solution.

Mots clés : Plans d'expériences optimaux ; Estimation de la matrice de trafic dans les réseaux IP ; Programmtion semi-définie ; Programmation conique du second ordre ; Optimisation sous-modulaire.

Optimal design of experiments with application to the inference of traffic matrices in large networks. *Second Order Cone Programming and Submodularity.*

Abstract: We approach the problem of optimizing the measurements in large IP networks, by using the theory of optimal experimental designs. This method gives raise to large scale optimization problems, for which we develop a resolution technique relying on Second Order Cone Programming (SOCP). The heart of our method is a rank reduction theorem in semidefinite programming. Some combinatorial problems –which arise when the goal is to find an optimal subset of the available experiments– are also studied.
The application to the inference of the traffic matrix in telecommunication networks is the object of the second part of this manuscript. We develop a method in which we optimize the estimation of several (randomly drawn) linear combinations of the traffic demands. This approach is compared to previous ones, and is fully evaluated by mean of simulations relying on real data. In particular, we handle some instances that were previously intractable.

Keywords: Optimal Experimental Design; Estimation of the traffic matrix in IP networks; Semidefinite programming; Second order cone programming; Submodular optimization.



MINES ParisTech

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY